UMA AVALIAÇÃO SOBRE CARACTERÍSTICAS ESPAÇO-TEMPORAIS LOCAIS BASEADAS EM INVARIANTES DE COR PARA RECONHECIMENTO DE AÇÕES

FILLIPE DIAS MOREIRA DE SOUZA

UMA AVALIAÇÃO SOBRE CARACTERÍSTICAS ESPAÇO-TEMPORAIS LOCAIS BASEADAS EM INVARIANTES DE COR PARA RECONHECIMENTO DE AÇÕES

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Arnaldo de Albuquerque Araújo
Coorientador: Guillermo Cámara Chávez

Belo Horizonte
Maio de 2011

FILLIPE DIAS MOREIRA DE SOUZA

AN EVALUATION ON COLOR INVARIANT BASED LOCAL SPATIOTEMPORAL FEATURES FOR ACTION RECOGNITION

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Arnaldo de Albuquerque Araújo
Co-advisor: Guillermo Cámara Chávez

Belo Horizonte
May 2011

Souza, Fillipe Dias Moreira de.

S729a Uma Avaliação sobre Características Espaço-Temporais Locais Baseadas em Invariantes de Cor para Reconhecimento de Ações / Fillipe Dias Moreira de Souza. — Belo Horizonte, 2011
xxiv, 79 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais. Departamento de Ciência da Computação.

Orientador: Arnaldo de Albuquerque Araújo.
Coorientador: Guillermo Cámara Chávez.

1. Computação - Teses. 2. Visão por computador - Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*84 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma avaliação sobre características espaço-temporais Baseadas em Invariantes
de Cor para Reconhecimento de Ações

FILLIPE DIAS MOREIRA DE SOUZA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Arnaldo de Albuquerque Araújo - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Guillermo Cámara Chávez - Co-orientador
Departamento de Computação - UFOP

Eduardo A. do Valle Jr.

Dr. Eduardo Alves do Valle Júnior
Bolsista de Pós-Doutorado-IC-UNICAMP

Prof. Nicolas Thome
Laboratoire d'Informatique de Paris 6 - UPMC

Prof. Adriano Alonso Veloso
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 13 de junho de 2011.

I dedicate this work to my parents, my great encouragers.

Acknowledgments

I end this work extremely grateful for the many years of support from my parents, professors, friends, and colleagues, during its development. Above all, I thank God for its subtle, yet strong, encouragement to make me do the things I have done to get to all my achievements. In this acknowledgment section, I would like to list a few names of people who have directly (say, more personally) influenced me and assisted me to finish my Master's dissertation. However, I would like to point out that I do not forget those not having their names mentioned and do not find their contribution less important. They are so many and have assisted me in so many different ways that I would have to write a book only for this purpose.

First, my thanks go to Prof. Arnaldo (UFMG, Brazil) for conferring me this opportunity to work with him as a Master student, in one of the best universities of my mother country, and introducing me to the exciting field of Image Processing and Computer Vision. I also thank him for allowing me to be co-supervised by great professionals of the field, Prof. Eduardo Valle (IC/UNICAMP, Brazil) and Prof. Guillermo Cámara Chávez (UFOP, Brazil). I owe both of them immense thanks for their direct assistance on many issues arisen during the development of my work, with respect to the theoretical as well as the practical matters. I also thank them for their valuable advices, suggestions, and endless discussions which permitted me to make refinements and, accordingly, improve my work. I also cannot leave behind my thanks to one person that has been continually helping me even after about three years not working together, Prof. Adriano Hoth (DCET/UESC, Brazil), my former advisor. In addition, I thank all of them for having guided me and instigated me to pursue a doctoral study outside the country.

I would like to thank the examination board's external members of my viva-voce defense for accepting the invitation to evaluate my work, for posing important questions on it, for spending their time in revising the written report and for giving suggestions to the final version. They are Prof. Adriano Veloso (DCC/UFMG, Brazil) and Prof. Nicolas Thome (LIP6/UPMC, France). Thank you everyone that watched my viva-voce presentation!

I thank Dr. Ivan Laptev (INRIA, France) for providing us with STIP's source code, which allowed us to focus only on modifying the code to deliver the proposals of the present work. Also, I thank Prof. Irving Biederman (USC, USA) for granting permission to reproduce a copyrighted picture from an excellent paper of his authorship. I also owe great thanks to a coworker of mine, and great friend, Mrs. Sandra Ávila (UFMG/UPMC, Brazil-France) for permitting the use of the Pornography dataset, previously conceived by her.

Especially, I would like to thank my parents, Fernando Antonio Moreira de Souza and Luciene Dias Sacramento de Souza, for all their invaluable incentive, assistance and words of encouragement throughout all those years of my engagement in higher education. Also, I thank my sister for her support even though well far away from me.

My thanks go specially to my girlfriend, Viviane Rodrigues da Silva, for her friendly support at all times of difficulty and happiness, and for constantly strengthening my willingness to do my best at finishing this work, with advices, special attention and love. I also thank her family (her parents, José Rodrigues da Silva and Maria Marly da Silva; her brothers, Juliene and Leonardo) and relatives (specially, her uncle Mauro and wife) for their support during my stay in Belo Horizonte and for the great times of entertainment and laughing.

Thank you all my friends and colleagues of the NPDI lab for giving me motivation and suggestions to get the most of this work. I also thank all of them for being so accommodating and friendly with me, and for the unforgettable happy hours. From those, I have to give special thanks to my friends Tiago Cunha, Sandra Ávila, Júlia, Marcelo Coelho, and colleagues Ana Paula, Elerson, André Taiar, Eduardo Gomes. Particularly, I am really grateful to Tiago for having given me an invaluable help when I first came to Belo Horizonte and more importantly for our friendship. Even though I only met Sandra when I came to Belo Horizonte, we have become great friends and I thank her for all days of work and dancing spent together!

I also thank Igor Calil and Bruno Teixeira for giving me the opportunity to work in their company as a software developer and have a worth professional experience.

Finally, I thank all my friends and colleagues that could not be here to see my accomplishment but were somewhere rooting for my success.

"Success means having the courage, the determination, and the will to become the person you believe you were meant to be."

(George Sheehan)

Resumo

Característica espaço-temporal local tem se demonstrado uma ferramenta poderosa para representar padrões não evidentes de objetos em movimento em cenas de vídeo, especialmente ações humanas. Reconhecimento de ações humanas é o foco principal para diversas aplicações, tais como indexação de vídeos, recuperação de vídeo baseada em conteúdo, resumo de vídeos, filtragem de conteúdo indesejável, classificação de filmes, para citar alguns.

De uma forma geral, detectores de pontos de interesse espaço-temporais contam somente com valores de luminância, ignorando cor. Além disto, a descrição de regiões de suporte são em sua maioria baseadas em histogramas de orientação por gradiente (para inferir descrição de forma) e fluxo óptico (para estimar movimentação aparente). Informação de cor tem sido em sua maioria ignorada durante os últimos anos de aprimoramento de técnicas para detecção e descrição de características locais no domínio espaço-tempo, apesar de ser comumente considerada um elemento importante para o entendimento de eventos.

Para reconhecimento de objetos e cenas em images estáticas, robustez a variações fotométricas foi alcançado através da descrição de regiões locais de pontos de interesse espaciais em termos das propriedades de invariância de cor. Em tal abordagem, robustez à geometria de iluminação, intensidade de iluminação e reflexão specular partiu do modelo de reflexão dicromática.

Neste contexto, o presente trabalho possue três contribuições principais. Primeiro, estendemos o detetor de esquinas espaço-temporais para incorporar informação de cor (utilizando o sistema de cor RGB-normalizado) na fase de deteção, o qual foi nomeado *ColorSTIP*. Segundo, consideramos o uso de histogramas de cor (baseado no canal *hue* ponderado pela saturação) para descrever regiões de suporte de pontos de interesse espaço-temporais, nomeando-o *HueSTIP*. Por fim, foi conduzida uma análise de desempenho criteriosa das extensões propostas para o reconhecimento de ações humanas em vídeos de cenários não controlados.

Palavras-chave: Características Espaço-Temporais, Invariantes de Cor, Pontos de Interesse Espaço-Temporais, Reconhecimento de Ações Humanas.

Abstract

Local spatiotemporal feature has been proved a powerful tool to represent latent patterns of moving objects in video scenes, especially human actions. Recognition of human actions is the principal focus for various applications, including video indexing, content-based video retrieval, video summarization, filtering of unwanted content, rating of movies, to name a few.

In general, spatiotemporal interest point detectors rely solely on luminance values, ignoring color. In addition to this, descriptions of the support regions are mostly based on histograms of gradient orientation (to infer shape description) and optical flow (to estimate motion appearance). Color information has been mostly overlooked during the last years of amelioration of techniques for detection and description of local features in the space-time domain, despite being usually considered an important element for the understanding of events.

For object and scene recognition in static images, robustness to photometric variations has been achieved by describing local regions of spatial interest points in terms of color invariance properties. In such approach, robustness to lighting geometry, illumination intensity and highlight was built on the dichromatic reflection model.

In this context, the present work holds three main contributions. First, we have extended the space-time corner detector (STIP) to incorporate color information (using the *normalized*-RGB color system) at the detection phase, which we have called the *ColorSTIP*. Secondly, we have considered the use of color histograms (based on the saturation-weighed hue channel) to describe support regions of spatiotemporal interest points, labeling it as *HueSTIP*. Finally, it was conducted a thorough analysis of performance of the proposed extensions for the human action recognition in videos on unconstrained (non controlled) scenarios.

Palavras-chave: Local Spatiotemporal Features, Color Invariants, Space-Time Interest Points, Recognition of Human Actions.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The creation of online social networks has greatly impacted the way people create, consume and share information, both in textual and visual media. Their use has fostered the creation of repositories where users work in cooperation to produce amounts of data that defy the imagination. For example, the number of video uploads to *YouTube* has recently reached about 35 hours per minute, the equivalent to 50,400 hours of video a day [YouTube, 2010].

The growth of multimedia data also stems from many other application domains, including surveillance systems, cultural heritage, geographic information systems, to name a few. Together with the increase in quantity, issues related to the complexity, quality, reliability and organization of multimedia data have also emerged. Information retrieval techniques have mostly addressed issues related to textual information (e.g., on the design of search engines for fast retrieval of relevant and reliable information, given very large corpora), whereas Computer Vision and Image Processing techniques have been employed to process visual data (e.g., for automating content-based video annotation and retrieval, video summarization, filtering of unwanted content).

Multimedia processing has now evolved enough to allow high-level semantic filters being used in practical applications. Demands of this nature come from many social contexts, including security, health, and quality of life. In the scope of security, much research has focused on i) automating the analysis of surveillance videos to help human operators detect suspicious and abnormal behaviors where safety and quietness must prevail [Siebel & Maybank, 2004], and ii) developing biometric systems for automated authentication of authorized persons. Examples regarding the quality of life include works for i) automated interpretation of sign language sentences to facilitate communication with deaf or mute people [Sarkar *et al.* , 2005], and ii) filtering of inappropriate content (e.g., pornography and violence scenes) for a specific audience [de Souza *et al.*

, 2010; Valle *et al.* , 2011].

Automated recognition of human actions and activities can favorably impact all such application contexts. Behavioral biometric system (i.e., based on human gait) were proved useful to uniquely identify people in [Sarkar *et al.* , 2005], and more convenient as it does not require the collaboration from users in order to collect biometric information. Interactive environments that react to human gestures are suggested to improve the user's experience in particular activities or places [Bobick *et al.* , 1999]. Being able to either predict or detect distinct anomalous behaviors from surveillance cameras could minimize the mistakes of human operators and thus enhance the security of public places, mainly in difficult conditions such as crowded environments. Correct categorization of human activities is a required feature for content-based video annotation and retrieval systems, as the recognition of specific actions can serve to imply the occurrence of certain events, thus providing some semantic power to such systems.

Motion events (or motion patterns) are advocated as the main source of information in order for human beings to interpret events and detect moving objects from the real-world [Wallach & O'Connell, 1953; Cutting *et al.* , 1978; Cedras & Shah, 1995]. Motivated by this fact, several works dedicated to put forward different ideas on how to detect and represent such motion patterns. Of the proposed concepts, motion patterns were modeled as trajectories of moving points, motion energy templates, global descriptors by optical flow, local spatiotemporal features, among many others.

Previously, the task of action recognition had been commonly addressed by the use of appearance-based methods with global description of motion. However, they can require costly preprocessing steps (e.g., segmentation and centering of the target object) before applying the method [Bobick & Davis, 2001] and, in addition, they can be more easily affected by undesirable episodes such as occlusions, viewpoint and lighting variations, and cluttered backgrounds typical of realistic scenes (for approaches not requiring a preprocessing step).

Methods based on local spatiotemporal features have gradually become the most popular for tackling the problem of action recognition, and almost completely supplanting their counterpart global methods. The main reasons for that lie in the fact that they are non-parametric approaches (meaning that they do not require the estimation of object model parameters) and are invariant to considerable changes of illumination and viewpoints, rotation, translation and occlusion (common problems on appearance-based global methods).

Local spatiotemporal features (or space-time interest points) are believed to provide important cues of motion patterns to represent the dynamic of moving image structures in scenes of videos. Several related techniques were proposed, evaluated and successfully applied to the automated recognition of human actions [Laptev, 2005; Dollár *et al.* , 2005; Willems *et al.* , 2008; Rapantzikos *et al.* , 2009]. Although there exists quite a few distinct local spatiotemporal feature detectors in the literature, they generally simply differ in terms of the design of the saliency measurement function. The saliency function holds proprieties that are inherited by the detected interest points and are responsible for making them discriminative (meaning that the elements have a unique representation) and robust to several variations (which can cause differences on the representation of the same element). Such points are expected to refer to patterns of motion by means of the description of their local regions (or support regions). The description of support regions in terms of histograms of oriented-gradients (to infer shape) and optic flow (to estimate apparent motion) is what define motion patterns in terms of local spatiotemporal features.

Motion events of real-world structures are intrinsic information for us to discern different dynamic episodes from real scenes, and therefore motion patterns are considered important cues for automated recognition of actions. Similarly, color information has a fundamental role in helping human beings to identify objects and interpret events from their surroundings. However, not much attention was drawn to either the design of space-time interest operators with color information or the use of color to describe those local points.

For static images, some works have proposed the use of color invariant properties to enrich the representation of image structures on automated object recognition and scene categorization [Gevers & Stokman, 2004; van de Weijer & Schmid, 2006; van de Sande *et al.* , 2010]. The authors in [van de Weijer & Schmid, 2006] extensively studied invariance properties of different color models based on the dichromatic reflection model. To improve the discriminative power and illumination invariance of local features to object recognition and image categorization, a set of color descriptors for spatial local features built on the work of [van de Weijer & Schmid, 2006] was proposed in [van de Sande *et al.* , 2010] for recognition of objects and scenes in static images. The distinctiveness of their color descriptors was evaluated experimentally and their invariant properties under illuminations changes were analyzed. They derived different color descriptors, including combinations with the intensity-based shape descriptor SIFT [Lowe, 2004].

This dissertation aims at merging the works in [van de Weijer & Schmid, 2006] and [Laptev, 2005], doing something similar to the work in [van de Sande *et al.* , 2010], but in the context of spatiotemporal features for action recognition.

1.1 Motivation

Most recently, there has been a great deal of interest in the development of new techniques for the extraction of local spatiotemporal patches for low-level representation of videos. Their main applications lie in automated interpretation of events such as recognition of human actions and categorization of events. Advances on techniques for representation of spatiotemporal patterns can possibly mean improvements on applications depending on automatic classification systems. Despite the efforts, for multiclass problems involving overlapping characteristics (that is, different classes containing similar visual patterns), classification rates are still low.

On the other hand, color information is commonly considered an important attribute to describe objects in the real world and an indispensable element to help us understand many complex situations. To the best of our knowledge, little attention was given to the use of color information to extract local spatiotemporal features from videos. This way, we found ourselves motivated to study and to contribute to the literature with this particular case.

1.2 Objectives

Based on the arguments given above, this work aimed to enrich local spatiotemporal features with photometric color invariants in both feature extraction phases, namely detection of local features in videos and their description. Our goal was to analyze the influence of color-based spatiotemporal features on specialized and generalized application contexts.

For the specialized case, we worked on two keenly-debated topics: pornography and violence. The exact meaning of "pornography" is a subject of controversy, but the concept has an important correlation with the one of nudity. As nudity tends to be typical of pornographic contents, skin color can be viewed as a key element to aggregate to the representation. On the violence context, the usefulness of color information is less obvious. In of spite of that, a few cases that may benefit from color information can be readily seen. For example, in professional fights the wrestlers are partially undressed (showing skin color) and bleeding of the participants is a not unusual event. Explosions as indicative of violence also have specific color patterns.

For the generalized case, we wanted to investigate if color information is a strong element to characterize more general types of events, where the importance of color is not so obvious *a priori* as in the pornographic scenario. In this regard, we found that this could only be fairly evaluated if tested in realistic scenes, where the settings are diverse in terms of color variations and happenings. To this end, we chose a dataset of Hollywood movies which fulfills this requirement and comprises a wide range of events. As events we refer to the different types of human actions in which the dataset is organized. This dataset is employed by many state-of-the-art evaluations.

In this context, our first hypothesis is that, for the events detected on the foreground (incited by motion of object parts), color may be a pivotal element in finely defining the parts actually playing a role to characterize the action. Following this reasoning, color will have the function of providing a more meaningful representation for the actions performed by the objects of interest. The second hypothesis stems from the idea that information about events in the background can contextualize, or restrict, what is happening on the foreground. Color information could then be added to the representation of the detected events in order to enhance their descriptions. Those premises were investigated in this dissertation.

It is known however that only the space-time interest points to represent motion patterns seem to be sufficient to improve the performance of classification problems depending on the motion pattern representation, as we have demonstrated in [de Souza *et al.* , 2010; Valle *et al.* , 2011]. For others, the key for success appears to be on how to fuse the elements of representation (color, optic flow, gradient orientations), as it was investigated in [Gevers *et al.* , 2006; Laptev *et al.* , 2008a].

There can be many ways of incorporating color information in the representation by local features, such as early and late fusion [Snoek *et al.* , 2005]. In early fusion, the low-level features are combined at the feature space level and the resultant artifact is used to infer the final concept. In late fusion, different feature-based views of the same data are individually learned and later combined to infer the final concept.

One first way attempted by us was to use the late fusion approach [Valle *et al.* , 2011], which we have found not to be an effective way. Others claim that color must be included through a more elaborate mechanism [Gevers *et al.* , 2006], so that its contribution is guaranteed. In this work, we incorporate color by two ways conforming the early fusion approach. One was to directly use the color channels of the normalized RBG system for the procedure of detecting space-time interest points. The other way was carried out by histogramming the color data (of the hue channel) from the surroundings of the spaiotemporal interest points, concatenating the color histogram to the other features.

Extending spatiotemporal features to include color data in their detection and description is an interesting and intuitive idea. But does color information provide a more meaningful representation to motion patterns? If so, how much is performance dependent on the task at hand, and even on each specific concept being detected? If yet true, are the improvements significant? Is it really worth the computational effort? Answering those questions is the main objective of this work.

More concretely, this dissertation investigates the attempt of incorporating photometric robustness to spatiotemporal features in terms of the zero-order color invariants' models proposed in [van de Weijer *et al.* , 2005; van de Weijer & Schmid, 2006], which were grounded by the dichromatic reflection model [Shafer, 1985].

1.2.1 Specific Objectives

1. Study the design of space-time interest point detectors and descriptors;
2. Investigate the state-of-the-art color invariant models in the literature;
3. Design and implement the alterations for the space-time interest point detector to incorporate color;
4. Design and implement the spatiotemporal color descriptor for the space-time interest points;
5. Test color-based local space-time interest points on generalized and specialized video datasets of human actions;
6. Analyze the experimental results under the assumption of different specific application contexts of current interest.

1.3 Contributions

The most important contribution of this work is the combination of the works van de Weijer & Schmid [2006] and Laptev [2005]. Color information is embedded into space-time interest point detector and, for the support regions, color is also considered for construction of the feature descriptors. Another contribution is a careful evaluation of this proposed extensions in comparison with the original STIP, on challenging datasets of real-world and complex scenes, investigating the cases of success and failure stemmed from the incorporation of color information.

1.4 Text Organization

In Chapter 2, a review on works on the main subjects of the present work is given, namely estimation of motion patterns and local spatiotemporal features. Details on how the local spatiotemporal features are found and how they gain meaning in terms of motion estimation and appearance description is also presented in Chapter 2. Chapter 3 expounds the fundamentals the target color invariants and describes the proposed extensions based on the invariants of the *normalized*-RGB color system and *hue* channel. Chapter 4 reports and analyzes the experimental results of the proposals of this work, on different datasets of videos involving scenes of human actions in realistic settings. At last, Chapter 5 concludes this dissertation.

Chapter 2

Related Work on Spatiotemporal Features

In this chapter, the concept of local feature is briefly introduced in Section 2.1 to give a general context of the main object of study in this dissertation. As motion events are one of the primary subjects in this work, an overall revision of the existing methods for motion estimation and interpretation will be outlined in Section 2.2. This section was almost completely based on the review given by [Laptev, 2004] and the survey in [Lopes *et al.* , 2010]. In addition, in our work motion patterns are represented as local spatiotemporal features, therefore the various methods proposed in the literature to localize features in space-time will be presented in Section 2.3 to give a first contextualization on the topic. Lastly, we present a few works related to the description of support regions for local spatiotemporal features, including methods closely related to the ones used in this dissertation.

2.1 A Brief Discussion on Local Features

There are numerous concepts associated to the term *local features* in image processing, but all of them can be summarized to the idea of locally gathered, invariant image patterns. (see Figure 2.1). Those patterns are associated with some image property, such as intensity, color or texture. Feature localization is essentially based on high intensity variations on local neighborhoods of those properties, and local features can be understood as primitive structures such as corners, edges, or blobs. Different types of features may have particular meanings depending on the application. For instance, in aerial images, detected edges often coincide with roads; while in automatic industrial product inspections blobs are compared with impurities [Tuytelaars & Mikolajczyk, 2008]. Furthermore, the name local feature can refer to local interest points or their descriptions. The description of the interest point serve as a compact and abstract representation of image patterns, which may take the form of histograms of gradient orientation [Kläser *et al.* , 2008; Lowe, 2004], optic flow [Lucas & Kanade, 1981] or color histograms [van de Weijer & Schmid, 2006; van de Sande *et al.* , 2010].

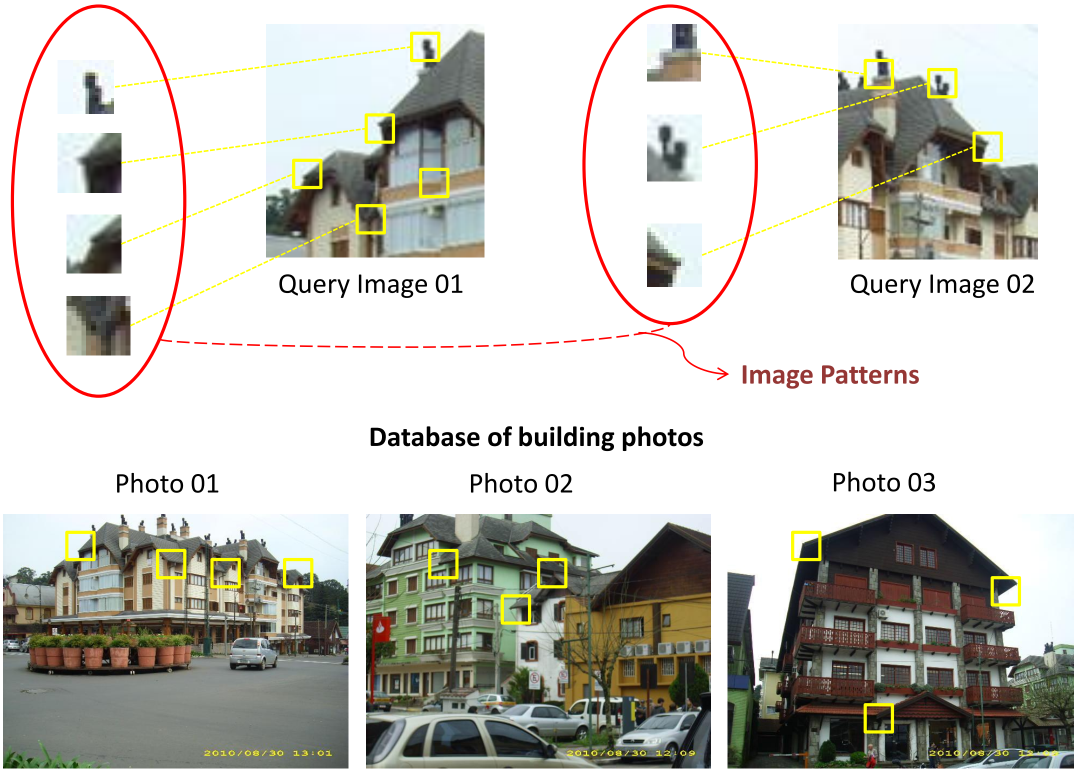

Figure 2.1. In this picture, the small squares represent image patterns. They are parts of image structures that compose specific objects, in this example, buildings. We expect that the same patterns are identified in images of similar content, i.e. which also display buildings.

Most successful methods proposed to localize local features in static images are based on intensity variations in local neighborhoods given by first- and second-order derivatives. They also employ smoothing with Gaussian windows of varying sizes to either reduce undesirable noises or to reproduce the same image content at different degrees of detail. However, many other strategies have been proposed different strategies to perform that same task. According to those strategies, local feature detectors can be organized into different categories, e.g., measurement of high-curvature points [Wang & Brady, 1995; Mokhtarian & Suomela, 1998], intensity based (e.g., measurements of intensity changes, and saliencies for locating interest points) [Beaudet, 1978; Moravec, 1981; Kitchen & Rosenfeld, 1982; Ullman & Sha'ashua, 1988; Forstner & Gulch, 1987; Harris & Stephens, 1988; Förstner, 1994], model-based [Guiducci, 1988; Giraudon & Deriche, 1991], color based [Würtz & Lourens, 1997; Montesinos *et al.* , 1998; van de Weijer *et al.* , 2005, 2006] to name a few. Owing to the scope of this work, we limit our discussion on techniques of interest points finding basis on corners.

According to Tuytelaars & Mikolajczyk [2008], local features own many properties that must be evaluated in order to verify the effectiveness and suitability of a single type. Assessing those properties qualifies feature types to distinct purposes. Repeatability and distinctiveness are conceivably the most important characteristics of a local feature type (especially in [Schmid *et al.* , 2000], these two characteristics are considered decisive to evaluate the power of an interest point detector). Regarding the former, from the several detected features of a group of images, the majority should be repeated in other sets of contextually related images, that is the chances of matching must be high. On what concerns the latter, the features are expected to have be distinctive but at the same time representative: this means that the same element (e.g., a corner of a particular object) should be given very similar values even if the object appears in different situations (illuminants, points of view, etc.), but different elements (e.g. corners of unrelated objects) should given very different values.

The use of local features to describe parts of image structures has not only been supported for its success in experimental research [Weber *et al.* , 2000; Agarwal & Roth, 2002; Fergus *et al.* , 2003; Lazebnik *et al.* , 2004], but has also been founded in the field of Cognitive Science [Biederman, 1987]. Biederman demonstrated that, on the human visual perception, objects are easily identified if information on their corners are retained, but not if corners are removed and only lines (edges) are kept (see Figure 2.2 for an illustrative example), stressing the importance of corners and junctions to visual recognition. Indeed, experimental results have proved the effectiveness of corner-based local features to applications of object recognition on several works(e.g., [Lowe, 1999; van de Weijer *et al.* , 2005; Laptev, 2006; van de Sande *et al.* , 2010]), including space-time corners for representation of video data [Laptev *et al.* , 2008b; Marszałek *et al.* , 2009].

2.2 Motion Interpretation of Moving Objects

In the context of video mining, a great deal of attention was drawn to the representation and estimation of motion patterns, as moving objects and image sequences are the

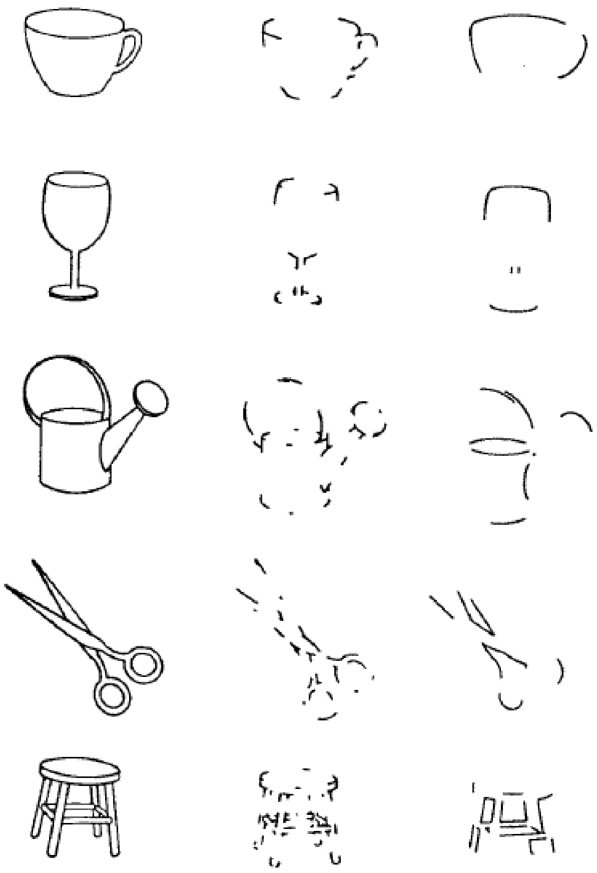

Figure 2.2. This picture outlines the idea that corners are important than lines to represent and easily identify objects. This picture is © American Psychological Association from [Biederman, 1987] (reproduced with permission).

main target and source of information, respectively, in order to understand events in videos. Accordingly, various methods were developed to satisfy such demand, which in [Laptev, 2004] were grouped into four types: structural methods, the use of motion templates for appearance-based methods, statistical appearance-based methods and motion interpretation based on events.

Structural methods Parameterized models defining the geometry of object parts and estimation of their mobility form the foundations of structural methods [Gavrila & Davis, 1995; Sidenbladh & Black, 2001; Sminchisescu & Triggs, 2003]. Such methods are directly related to motion capture techniques and objects in the scene are supposed to behave in accordance with specifically parameterized models. In such cases, different sets of parameters match distinct events, and parameters are associated with the location of object parts.

Because of the high number of parameters to be estimated, constraints can be introduced in order to attenuate the burden of parameter search (structural methods are typically constrained to a number of assumptions, such as static background and definite scale). Even though commonly approached with three-dimensional reconstruction of the image structures of interest, structural methods have alternatively used image features for tracking and analyzing motion trajectories, for example in [Bregler, 1997; Isard & Blake, 1998; Song *et al.* , 2003].

In Lopes *et al.* [2010], related structural methods are categorized as parametric object methods. Parametric object model methods can comprise two more subgroups, namely internal models, and trajectory models. Works based on the former's direction define the structural internal models of the objects of interest and then adapt the computed parameters to the visual information of the scene [Gavrila & Davis, 1995; Sidenbladh & Black, 2001; Sminchisescu & Triggs, 2003; Yilmaz & Shah, 2005; Gaitanis *et al.* , 2006; Filipovych & Ribeiro, 2008], whereas on the latter information of the object internal states is ignored and it counts solely on tracking of the object location over the frame sequence [Bobick & Davis, 2001; Abdelkader *et al.* , 2008; Hu *et al.* , 2008].

For relying on explicit locations of parts, structural methods have a favourable position for applications of human-computer interfaces and object animation, whereas unreliable to interpretation of events in realistic scenes [Laptev, 2004].

Motion templates for appearance-based method In this method, motion is inferred by analyzing appearance changes of objects over the sequence of frames. Bobick & Davis [2001] presented a seminal work in this branch. They proposed the construction of 2D temporal motion templates to build action recognition systems. Those templates were known as Motion Energy Image (MEI) and Motion History Image (MHI). In the MEI model, a motion template of one specific action was represented by a binary image, where the region filled with high values (white pixels) accounted for the motion area. For the MHI model, a motion template of an action was a gray-scale image, in which the brighter a specific range of pixels, the more recent was motion at that area.

This formed the ground to more other related work, such as in [Efros *et al.* , 2003; Hu *et al.* , 2009]. The works in [Black *et al.* , 1998; Hoey & Little, 2003; Efros *et al.* , 2003] focused on describing and recognizing motion events in terms of optic flow in attempt of preventing precise spatial segmentation. In addition, 3D templates in terms of 3D volumes served in [Gorelick *et al.* , 2007] as a more attractive solution for representing actions in image sequences, yet based on silhouettes, since they were expected to carry static and dynamic information at the same time. Several other works followed this idea, namely [Mokhber *et al.* , 2008; Cuntoor, 2006; Fathi & Mori, 2008].

The main advantage of using motion template based approaches rather than structural methods comes from the fact that a less quantity of parameters is required, which possibly helps to diminish equivocal matchings. Following the organization proposed in Lopes *et al.* [2010], motion templates fit in the class of implicit object models.

Statistical appearance-based methods To elude the problem of finding explicit correspondence between points of object templates and object structures in unknown data, statistical (or non-parametric) methods combined with position-independent image measurements were regarded as a tangible alternative to simplify the problem to only object motion pattern classification. This way, the problem is reduced to only discover which types of motion patterns exist in the scene, instead of also having to find their location. To deal with this, various works have proposed different non-parametric methods in the literature.

To start with, motion patterns from flowing water, waving cloth and fluttering paper were classified by using first- and second-order statistics of normalized spatiotemporal gradients in [Polana & Nelson, 1997]. An extension of this work was introduced by Chomat & Crowley [1999], in which recognition of human gestures and activities was tackled by means of histograms of spatiotemporal filter responses. Clustering of events having repetitive activities and human action recognition were performed by employing marginalized histograms of multiscale spatiotemporal gradients by Zelnik-Manor & Irani [2001]. Other works on statistical methods can be found in [Fablet *et al.* , 2002; Doretto *et al.* , 2003]. Despite the promising results presented by those statistical methods, spatial segmentation remains a drawback, since implicit correspondence is still required. This means that such methods are sensitive, for instance, to occlusions, motions of cameras, and motion in the setting (background).

On what concerns statistical appearance-based methods, the experimentally evaluated motion descriptors that demonstrated the highest performance in [Laptev, 2004] (on a controlled-scenario dataset of human actions) were used in this dissertation to characterize the detected local space-time features. The descriptors were defined in terms of spatiotemporal gradient and optic flow position-dependent histograms, which were computed separately and joint afterwards. Those methods will be utterly described in Section 2.6.

Motion interpretation based on event patterns The explicit correspondence of points required by some of the methods mentioned above is resolved by only considering spatial variations. Temporal variations, however, can be helpful to improve matching and localization of representative motion patterns (regarded as *motion events*). Early works on this, namely in [Rubin & Richards, 1985; Engel & Rubin, 1986], qualified motion events as *motion boundaries*, being interpreted as smooth starts, smooth stops, pauses, impulse starts and impulse stops. This idea stuck to considerations on kinematics, comparing motion boundaries with force discontinuities of real situations.

The related work in [Gould & Shah, 1989] created a scheme of trajectory-based motion representation referred to as TPS (Trajectory Primal Sketch). The TPS representation encodes a set of trajectory primitives defining motion events that should coincide with discontinuities in speed, direction and acceleration. That proposal aimed to make feasible the identification of object structures by only using their supposedly unique trajectory signatures.

Brand [1997] stressed the mechanism of retrieving primitive events from image sequences, claiming the importance of visual events characterized by splits and unifications. However, to detect events, his method requires a previous step of spatial segmentation which isolates the image structures in movement. This clearly happens to be a drawback when dealing with videos displaying too heterogeneous settings.

In another approach closely following the principles defended in [Rubin & Richards, 1985; Engel & Rubin, 1986], global descriptors of optic flow were used to describe videos [Rui & Anandan, 2000]; analyzing the flow over the sequence of frames and determining that substantial changes in flow indicated motion boundaries.

To alleviate the drawbacks inherent to the early methods of event-based motion interpretation, such as the computation of trajectories, spatial segmentation, and global image measurements, Laptev [2005] proposed a scheme to represent object dynamics in terms of local information. By using local information, the undesirable effects from complex scenes caused by occlusions and image clutter can be avoided, or minimized, yet escaping the necessity of solving the correspondence problem over the sequence of frames required by methods depending upon trajectory analysis.

For extracting local motion patterns from image sequences, his method was built on the combined edge-corner detector proposed by Harris & Stephens [1988] (well-known as Harris corner detector), extending it to the temporal dimension. This is a 3D version of the Harris corner detector that counts on localizing significant intensity changes not only spatially, but also over time. In fact, this method, the Spatiotemporal Interest Point (STIP) detector, is more exactly an extension of the scale-invariant Harris corner detector proposed by Mikolajczyk & Schmid [2002] (referred to as Harris-Laplace corner detector). This multiscale version became invariant to scale by applying the Laplacian of Gaussian (LoG) operator on the automatic selection of the different scales [Mikolajczyk & Schmid, 2002]. This local space-time interest point detector is the basis for the present dissertation and will be further detailed in Section 2.5.

2.3 Local Spatiotemporal Feature Detectors

In this section, we review on the several existing spatiotemporal interest point detectors. Algorithms of interest point detection have their core contribution on the design of a saliency measurement function, such that the *interestingness* of local regions is signalized by characteristic values produced by such function. The first extension of a spatial interest point detector to simultaneously take into account the variations over time was proposed by Laptev & Lindeberg [2003], which, as mentioned earlier, was built on the multiscale Harris corner detector. The response function of Harris is expressed by a simple operation involving the determinant $\det(\Lambda)$ and the trace $\text{trace}(\Lambda)$ of an auto-correlation matrix Λ (a second-moment matrix), such that when $H = \det(\Lambda) - k.\text{trace}(\Lambda)$ is maximum a corner has been found.

$$\Lambda = \begin{pmatrix} L_xL_x & L_xL_y \\ L_xL_y & L_yL_y \end{pmatrix}, \tag{2.1}$$

where L_i is the Gaussian derivative of the image at dimension i (where $i \in \{x, y\}$).

The local maxima triggering measure H is intimately related to the eigenvalues (λ_1 and λ_2) of matrix Λ. In order to avoid their computation, the corner response is given as function of the determinant and trace, which are in turn closely related to the eigenvalues. Laptev & Lindeberg [2003] extends this matrix to the temporal domain by adding the derivative of L in the time direction t, so Λ coarsely becomes

$$\Lambda = \begin{pmatrix} L_x L_x & L_x L_y & L_x L_t \\ L_x L_y & L_y L_y & L_y L_t \\ L_x L_t & L_y L_t & L_t L_t \end{pmatrix}. \tag{2.2}$$

Followed by the promising achievements in human action recognition presented in [Schuldt *et al.* , 2004] using the spatiotemporal extension of Harris detector, many other related work in the literature arose. An investigation on the use of STIP to other contexts, namely rodent behavior analysis and facial expressions, revealed its weakness [Dollár *et al.* , 2005]. This contributed to the proposal of another detector of spatiotemporal features by Dollár *et al.* based on Gabor filters, called *cuboid* detector.

Cuboids are localized when the response function $R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$ produces high values in the neighborhood of a point (x, y). One-dimensional Gabor filters,

$$h_{ev}(t; \tau, \omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2} \tag{2.3}$$

and

$$h_{od}(t; \tau, \omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2}, \tag{2.4}$$

are both temporally applied to a sequence of smoothed images, but this detector is variant to scale. The Gabor filter's Equations 2.3 and 2.4, $\omega = 4/\tau$ and τ is the temporal scale. In addition, the images are smoothed by two-dimensional Gaussian kernels $g(x, y; \sigma)$, where σ is the spatial smoothing scale.

Subsequently, built on the concept of salient points by Kadir & Brady [2001], Oikonomopoulos *et al.* [2006] proposed an extension to consider activity variation in a cylindrical spatiotemporal neighborhoods defined by distinct radii s and temporal depths d (parameters which determine a specific scale). In addition, intensity information is determined by convolution with a Gaussian window of first-order derivatives in the temporal domain, such that, given an image sequence $I_0(x, y, t)$, the input signal is $I(x, y, t) = g_t * I_0(x, y, t)$.

Given a cylindrical region $N_{cl}(\vec{s}, \vec{v})$ (that is a set of pixel values) centered at $\vec{v} = (x, y, t)$ in scale $\vec{s} = (s, d)$, the saliency metric $y_D(\vec{s}, \vec{v})$ measures the alteration levels of the information content encompassed by $N_{cl}(\vec{s}, \vec{v})$ (where D is the set of possible signal values). Basically, it consists of the amount of entropy $H_D(\vec{s}, \vec{v})$ measured in a specific spatiotemporal region weighed by its degree of relevance $W_D(\vec{s}, \vec{v})$ at a certain scale $\vec{s}$ (see Equations 2.5 2.6 2.7, where p_D is the probability density of the signal histogram, q is the signal value and D is the range values originated from $I(x, y, t)$).

$$y_D(\vec{s}, \vec{v}) = H_D(\vec{s}, \vec{v}) W_D(\vec{s}, \vec{v}), \tag{2.5}$$

$$H_D(\vec{s}, \vec{v}) = -\int_{q \in D} p_D(\vec{s}, \vec{v}) \log_2 p_D(\vec{s}, \vec{v}) dq, \tag{2.6}$$

and

$$W_D(\vec{s}, \vec{v}) = s \int_{q \in D} |\frac{\partial}{\partial s} p_D(\vec{s}, \vec{v})| dq + d \int_{q \in D} |\frac{\partial}{\partial d} p_D(\vec{s}, \vec{v})| dq. \tag{2.7}$$

Until that recent, only approaches involving filtering and entropy computation of local regions had been proposed. A very dissimilar scheme for detecting spatiotemporal interest points was then presented by Wong & Cipolla [2007], arguing that global information could be as effective as local regions to identify discriminative and compact representations of object parts. In that work, the authors developed a novel algorithm for detecting interest points over space and time motivated by the work on analysis of dynamic textures.

In the sequel, Willems *et al.* [2008] generalized the Hessian matrix Hes to the spatiotemporal domain (see Equation 2.8), where the matrix elements are second-order partial derivatives $(L_{xx}, L_{xy}, ..., L_{tt})$. The authors also proved the feasibility of localizing features simultaneously in both the spatiotemporal and multiscale domains. Unlike the work in Laptev [2005] that requires an iterative procedure to automatically select characteristic scales, their multiscale approach is a non-iterative scheme and because of this helps to considerably reduce the computational burden required for attaining scale-invariance.

3D convolutions approximated using box-filters applied to integral videos (extension of the integral image concept to image sequences) justify the computational efficiency of the algorithm. Local extrema responds to high absolute value of the matrix determinant $S = |\det(Hes)|$ (which can be seen as an extension from [Beaudet, 1978]). This criteria does not assure that all eigenvalues have the same sign. In such case, the authors suggest that a post-processing procedure could be applied to discard the saddle points by only checking the signs (claiming that this is not a serious problem provided that points are reliably detected, that is, have high repeatability).

$$Hes = \begin{pmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{xy} & L_{yy} & L_{yt} \\ L_{xt} & L_{yt} & L_{tt} \end{pmatrix}. \tag{2.8}$$

Rapantzikos *et al.* [2009] approach the problem of detecting spatiotemporal features in a peculiar way yet following the same principal idea of saliency measure. Intensity, motion and color information all contribute to exploit the latent patterns in space and time by means of an energy minimization formula, which is essentially composed by two primary terms, namely the data term and the smoothness term. The incorporation of those three types of data source is given by the smoothness component. Such component happens to be an interaction between the distinct types of features incorporated, such that each type yields independent components referred as intensity conspicuity, color conspicuity (which is based on the opponent color theory) and orientation conspicuity (which involves the use of steerable filter for determining the orientations). A thorough description of this method can be found in [Rapantzikos *et al.* , 2009].

2.4 Spatiotemporal Descriptors for Action Recognition

In this section, we review on the several descriptors used in the literature to describe event dynamics occurring in image sequences of human action datasets. Laptev & Lindeberg [2004] evaluated descriptions of local space-time regions by means of i) single and multiscale N-jets (an N-jets descriptor is defined as a vector of spatiotemporal derivatives up to N order), ii) histograms of spatiotemporal gradients and optic flow with dimensionality reduced by the Principal Component Analysis (PCA) technique, and iii) position-dependent and position-independent histograms of spatiotemporal gradients and optic flow, such that the combination holding the best performance was given by the position-dependent histograms of spatiotemporal gradients and optic flow (which are going to be later discussed in details).

In Dollár *et al.* [2005], cuboid descriptors were based on three types of transformations, namely i) normalized pixel values, ii) brightness gradients, and iii) windowed optic flow. All of those method had dimensionality reduced by the PCA technique, but the best performance was achieved by the flattened gradient descriptor, which in turn was a generalization of the PCA-SIFT descriptor presented in Ke & Sukthankar [2004].

Built on the success of the Lowe's 2D SIFT descriptor, Scovanner *et al.* [2007] proposed the 3-dimensional SIFT descriptor, especially to tackle the problem on action recognition. Similar to the 2D SIFT, the generalization to space-time computes gradient magnitude and orientation at each pixel of the local region defined by the spatiotemporal interest point. However, to better capture the spatiotemporal nature of space-time patterns, the orientation is represented by a pair of angles computed as $\theta(x, y, t) = \tan^{-1}(L_y/L_x)$ and $\phi(x, y, t) = \tan^{-1}(L_t/\sqrt{L_x^2 + L_y^2})$, in which L denotes the LoG approximation. The descriptor is a histogram comprised of bins uniquely identified by the pairs (θ, ϕ). Each pixel in local neighborhood will be assigned a pair (θ, ϕ), and the histogram distribution of the support region is quantized by the magnitudes at such pixels.

Another approach for building spatiotemporal gradient descriptors was introduced in Kläser *et al.* [2008]. The histogram construction was based on regular polyhedrons in which the histogram bins are represented by their faces. Yet, it counted on integral video concept (a generalization of the work in [Viola & Jones, 2001]) to provide a more computationally efficient description algorithm. Lopes *et al.* [2009] proposed detecting and describing 2D features from which they called spatiotemporal frames (given that a video is represented by 3 dimensions (x, y, and t), local features were extracted from spatiotemporal frames xy, xt and yt). They demonstrated that this plain 2D alternative of representing dynamic information could reach comparable performance with more sophisticated 3D detectors of motion patterns on human action dataset of controlled scenario.

To account for motion patterns, this dissertation found basis on the work of Laptev [2005], which represents latent motion patterns in terms of local spatiotemporal interest points. The proposed descriptors that demonstrated the best performance in describing the points were also considered for use in our work. In the sequel, we provide the reader with a formal discussion on the theoretical foundations of the spatiotemporal interest point detector and their descriptors before we present the proposed modifications in Chapter 3.

2.5 Detecting Spatiotemporal Corners

Now we broaden the discussion about the detection of space-time interest points (STIP) that we have drafted in Section 2.3. First, we present the theory of interest point detection in the spatial domain. Next, we extend of the method to the temporal dimension.

The detection of interest points (as corners) in the spatial domain (e.g., static images) can be described as follows. The linear scale-space representation of an image can be mathematically defined as $L^{sp} : R^2 \times R_+ \mapsto R$, which is the convolution of f^{sp} with g^{sp}, where $f^{sp} : R^2 \mapsto R$ represents a simple model of an image and g^{sp} is the Gaussian kernel of variance σ_l^2. Then,

$$L^{sp}(x, y; \sigma_l^2) = g^{sp}(x, y; \sigma_l^2) * f^{sp}(x, y), \tag{2.9}$$

and

$$g^{sp}(x, y; \sigma_l^2) = \frac{1}{2\pi\sigma_l^2} \exp(-(x^2 + y^2)/2\sigma_l^2). \tag{2.10}$$

Localizing interest points means to find strong variations of image intensities along the two directions of the image. To determine those local regions, the second moment matrix is integrated over a Gaussian window having variance σ_i^2, for different scales of observation σ_l^2, which is written as the equation:

$$\begin{aligned} \mu^{sp}(.; \sigma_l^2, \sigma_i^2) &= g^{sp}(.; \sigma_i^2) * ((\nabla L(.; \sigma_l^2))(\nabla L(.; \sigma_l^2))^T) \\ &= g^{sp}(.; \sigma_i^2) * \begin{pmatrix} (L_x^{sp})^2 & L_x^{sp} L_y^{sp} \\ L_x^{sp} L_y^{sp} & (L_y^{sp})^2. \end{pmatrix} \end{aligned} \tag{2.11}$$

The descriptors of variations along the dimensions of f^{sp} are the eigenvalues of Equation 2.11: λ_1 and λ_2, with $\lambda_1 \leq \lambda_2$. Higher values of those eigenvalues is a sign of interest point and generally leads to positive local maxima of the Harris corner function, provided that the ratio $\alpha = \lambda_2/\lambda_1$ is high and satisfies the constraint $k \leq \alpha/(1+\alpha)^2$:

$$\begin{aligned} H^{sp} &= \det(\mu^{sp}) - k.\text{trace}^2(\mu^{sp}) \\ &= \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2 \end{aligned} \tag{2.12}$$

.

Analogously, the procedure to detect interest points in the scape-time domain is derived by rewriting the equations to consider the temporal dimension. Thus, having an image sequence modeled as $f : R^2 \times R \mapsto R$, its linear representation becomes $L : R^2 \times R \times R_+^2 \mapsto R$, but over two independent variances σ_l^2 (spatial) and τ_l^2 (temporal) using an anisotropic Gaussian kernel $g(.;\sigma_l^2,\tau_l^2)$. Therefore, the complete set of equations for detecting interest points described in [Laptev & Lindeberg, 2003] is the following.

$$L(.;\sigma_l^2,\tau_l^2) = g(.;\sigma_l^2,\tau_l^2) * f(.), \tag{2.13}$$

$$\begin{aligned} g(x,y,t;\sigma_l^2,\tau_l^2) &= \frac{1}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}} \\ &\times \exp(-(x^2+y^2)/2\sigma_l^2 - t^2/\tau_l^2), \end{aligned} \tag{2.14}$$

$$\mu = g(.;\sigma_i^2,\tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix}, \tag{2.15}$$

and

$$\begin{aligned} H &= \det(\mu) - k.\text{trace}^3(\mu) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3, \end{aligned} \tag{2.16}$$

restricted to $H \geq 0$, with $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$, and subject to $k \leq \alpha\beta/(1+\alpha+\beta)^3$

2.6 The STIP Descriptors

Several kinds of descriptor methods were tested to extract descriptive information of shape and motion from local neighborhoods of the space-time interest points detected by STIP. Position-dependent histograms of gradient orientation (HoG) and histograms of optic flow (HoF) have shown appealing results in comparison to the others [Laptev & Lindeberg, 2004]. The combination of those histograms, i.e. the concatenation of them into one single descriptor vector, was proved the configuration that best contributed to increased rates of classification performance for action recognition.

The construction of a descriptor may be accomplished by three distinct manners [Dollár *et al.* , 2005]. The most simple way involves flattening the local region of pixels into a vector, but sensitive to small disturbances. Histogramming the values from the local patch is possibly a second alternative, which cannot preserve positional information but claims robustness to disturbances. The third concerns with the segmentation of the local patch into smaller areas, and posterior computation of local histograms in those places. The local histograms are concatenated to form a single one. This last approach preserves location and robustness to perturbations, which is the one used in Lowe [2004]; Dollár *et al.* [2005]; Laptev & Pérez [2007]. Furthermore, the STIP's descriptors are computed at different scales, where each scale highlights a level of detail to which the same scene may be observed. The spatial scales σ can assume the values 4 and 8, while the temporal scales used are 2 and 4.

2.6.1 Oriented-Gradient Histogram

The use of gradient directions as local features has been motivated by its capability of inferring shape information, which was attested by its success in object recognition [Dalal & Triggs, 2005; Laptev, 2006; Lowe, 1999]. In this approach, every pixel of a selected local patch has its gradient orientation computed with respect to its neighborhood. The local patch encompassing the detected interest point is divided into smaller blocks corresponding to different local positions in the local patch (which makes it a position-dependent histogram). From each sub-block a histogram of gradient orientation will be originated through a quantization process of gradient orientation bins weighed by the amounts of magnitude. The magnitude m and orientation θ can be calculated as

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y-1) - L(x,y-1))} \tag{2.17}$$

and

$$\theta(x,y) = \arctan \frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}. \tag{2.18}$$

In this dissertation, a spatiotemporal patch was sliced into 3 x 3 x 2 blocks, such that 4-bin gradient orientation histograms were computed to every block, resulting in histograms of 72 bins (a 72-dimensional feature vector). To visualize an illustration of this process, see Figure 2.3.

2.6.2 Descriptor of Optical-Flow

Optical flow accounts for the estimation of apparent motion of objects, through a field of velocity vectors, in one image sequence. It infers information about motion of objects by computing the distribution of apparent velocity from changes in intensity

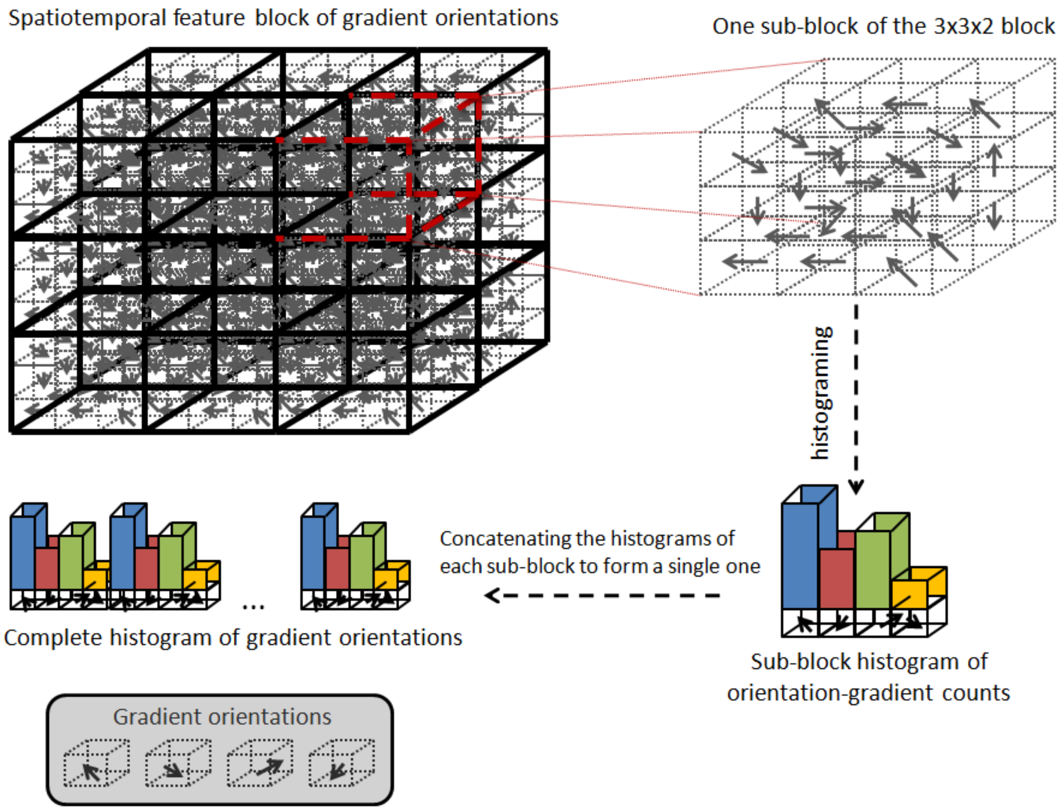
Spatiotemporal feature block of gradient orientations
One sub-block of the 3x3x2 block
histograming
Concatenating the histograms of each sub-block to form a single one
...
Complete histogram of gradient orientations
Sub-block histogram of orientation-gradient counts
Gradient orientations

Figure 2.3. This diagram depicts the mechanism for construction of oriented-gradient histograms from spatiotemporal features.

patterns. Apparent motion can be understood as what results from what is perceived by the observer when objects are moving in the scene, or the reverse. That means that motion via optic flow happens either by the movement of structures in the scene or by motion of the observer (e.g., a camera).

Different approaches have been proposed to compute optic flow [Barron *et al.* , 1994]. In [Barron *et al.* , 1994], they could be categorized into the region-based [Anandan, 1989; Singh & Allen, 1992], differential-based [Lucas & Kanade, 1981] [Horn & Schunck, 1980] [Nagel, 1982; Uras *et al.* , 1988], energy-based [Heeger, 1987; David J., 1988], and phase-based [Waxmann *et al.* , 1998; Fleet & Jepson, 1990] methods, to name a few. From the myriad of proposals, the Lucas-Kanade gradient-based technique was used in this work, since it makes part of the original implementation of STIP's motion descriptor. Supposing an intensity image I, one important premise stated by gradient-based approaches is that each pixel brightness $I(x, y, t)$ has a constant value after a short period of time δt, thus only suffering shift of spatial location $(\delta x, \delta y)$. This way,

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t), \tag{2.19}$$

where $(\delta x, \delta y)$ denotes the displacement of the pixels in both spatial directions x and y after the time interval δt. The displaced image can then be neatly approximated by the first-order Taylor series as follows

$$I(x+\delta x, y+\delta y, t+\delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t. \tag{2.20}$$

Then, it follows that

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = 0 \tag{2.21}$$

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t}\frac{\delta t}{\delta t} = 0, \tag{2.22}$$

which could have also been reached by using the chain rule to derive I with relation to t, where $\nabla I \equiv (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y})$ and $\frac{\partial I}{\partial t}$ are the image sequence's spatial and temporal derivatives, and the differentiations $u = \frac{\delta x}{\delta t}$ and $v = \frac{\delta y}{\delta t}$ denote the desirable velocity vectors (comprising the 2D velocity vector $\vec{u} = (u, v)$). The Equation 2.6.2 is known as the constraint equation, and becomes

$$\nabla I.\vec{u} = -\frac{\partial I}{\partial t}, \tag{2.23}$$

which is one equation with two unknowns, and thus cannot be directly solved unless an additional set of constraint equations is available (the aperture problem). In order to solve this problem and to be able to calculate the components of the velocity field, Lucas & Kanade use the local neighborhood of the pixel at a certain position, with a definite size $N \times N$, generating one restriction equation to each pixel position of such region. Thus, the computation of the velocity vectors at a certain position is supported by N^2 equations,

$$\begin{aligned}
\frac{\partial I}{\partial x_1}u + \frac{\partial I}{\partial y_1}v &= -\frac{\partial I}{\partial t_1}\\
\frac{\partial I}{\partial x_2}u + \frac{\partial I}{\partial y_2}v &= -\frac{\partial I}{\partial t_2}\\
\frac{\partial I}{\partial x_3}u + \frac{\partial I}{\partial y_3}v &= -\frac{\partial I}{\partial t_3}\\
&\vdots\\
\frac{\partial I}{\partial x_{N^2}}u + \frac{\partial I}{\partial y_{N^2}}v &= -\frac{\partial I}{\partial t_{N^2}}
\end{aligned} \tag{2.24}$$

which turns out to be an overdetermined system,

$$\begin{bmatrix} \frac{\partial I}{\partial x_1} & \frac{\partial I}{\partial y_1} \\ \frac{\partial I}{\partial x_2} & \frac{\partial I}{\partial y_2} \\ \vdots & \vdots \\ \frac{\partial I}{\partial x_{N^2}} & \frac{\partial I}{\partial y_{N^2}} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -\frac{\partial I}{\partial t_1} \\ -\frac{\partial I}{\partial t_2} \\ \vdots \\ -\frac{\partial I}{\partial t_{N^2}} \end{bmatrix} \tag{2.25}$$

or

$$W\vec{u} = -b. \tag{2.26}$$

To solve the overdetermined system, Lucas & Kanade apply the least squares method, then

$$\begin{aligned} W^T W\vec{u} &= W^T(-b) \\ \vec{u} &= (W^T W)^{-1} W^T(-b), \end{aligned} \tag{2.27}$$

which leads to

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i^{N^2} \frac{\partial I}{\partial x_i}^2 & \sum_i^{N^2} \frac{\partial I}{\partial x_i}\frac{\partial I}{\partial y_i} \\ \sum_i^{N^2} \frac{\partial I}{\partial x_i}\frac{\partial I}{\partial y_i} & \sum_i^{N^2} \frac{\partial I}{\partial y_i}^2 \end{bmatrix} \begin{bmatrix} -\sum_i^{N^2} \frac{\partial I}{\partial x_i}\frac{\partial I}{\partial t_i} \\ -\sum_i^{N^2} \frac{\partial I}{\partial y_i}\frac{\partial I}{\partial t_i} \end{bmatrix} \tag{2.28}$$

.

To construct the histograms of optical flow, the same mechanism used to build histograms of gradient orientation was used, except for having the quantization process based on 5 bins, with four bins making reference to motion in distinct directions, and the last bin representing the absence of motion. As the spatiotemporal patch comprises of $3 \times 3 \times 2$ blocks, the final histogram of optical flow contains 90 bins.

2.7 Summary

In this chapter, we have discussed the fundamental theory of the motion pattern detector used in this work. The detector is the state-of-the-art space-time interest point operator STIP, which is an adaptation of the Harris-Laplace corner detector to the space-time domain. We also briefly presented both feature types used to characterize the neighborhood of the detected interest points: optical flow and gradient orientation. The final feature descriptor of one space-time interest point is resulted from the concatenation of both histograms of features. This original version of STIP will have its performance further compared on the problem of action recognition with different adaptations of it based on color information.

Chapter 3

The Proposed Color Invariant Based Methods

As stated earlier, our proposal aims at investigating if color invariants incorporated to the original STIP could boost performance rates on the problem of classification of human actions in realistic settings. The color models applied here are the models of zero-order color invariants derived in [van de Weijer & Schmid, 2006], namely the *normalized*-RGB system (hereinafter referred to as *rgb*, to follow the convention adopted in [van de Weijer & Schmid, 2006]) and the *hue* channel from the HSI system. At first glance, those choices may appear simplistic and even arbitrary — therefore this chapter is concerned with a careful explanation of the basis of the used color models, comprising of a discussion on the dichromatic reflection model [Shafer, 1985] and how the color invariants are derived from it. At the end of this chapter, we show the revisions that were made to the processes of detection and description of space-time interest points, so that the family of color-based STIPs was generated.

3.1 Color Models

There exists a myriad of color models (or color systems) being used, and each originally designed to fulfill various purposes. Essentially, a color system is created to standardize the definition of color in a particular application. The most common of them is the RBG system (where RGB stands for Red, Green, Blue), which is a scheme of color estimation emulating the functioning of human vision color perception in a finite space. Each dimension of this orthogonal 3D space accounts for a wavelength light to which our visual photo-sensitivity molecules are most reactive. To form other colors using those primaries, the beams of light in each wavelength must be superposed, meaning that the RGB system is additive. Color monitors and a wide range of cameras are understructured by the RGB system to display and capture colorful visual information.

In an 8-bit representation, the axes values range from 0 to 255, where 0 indicates the absence of the primary color, 255 corresponds to its full intensity, and the middle interval samples its different proportions. The cube diagonal having as extremes (0,0,0) and (255,255,255) symbolizes the gray-scale axis (in which color is undefined), such that points on that line have equal amounts of R, G and B; for instance, the mid-gray is at position (128,128,128). Thus, to identify any color inside the cube, it is just needed to project the three components' amounts in the space. A geometric illustration of this discussion, as well as a colorful visualization of the RGB space is depicted in Figure 3.1.

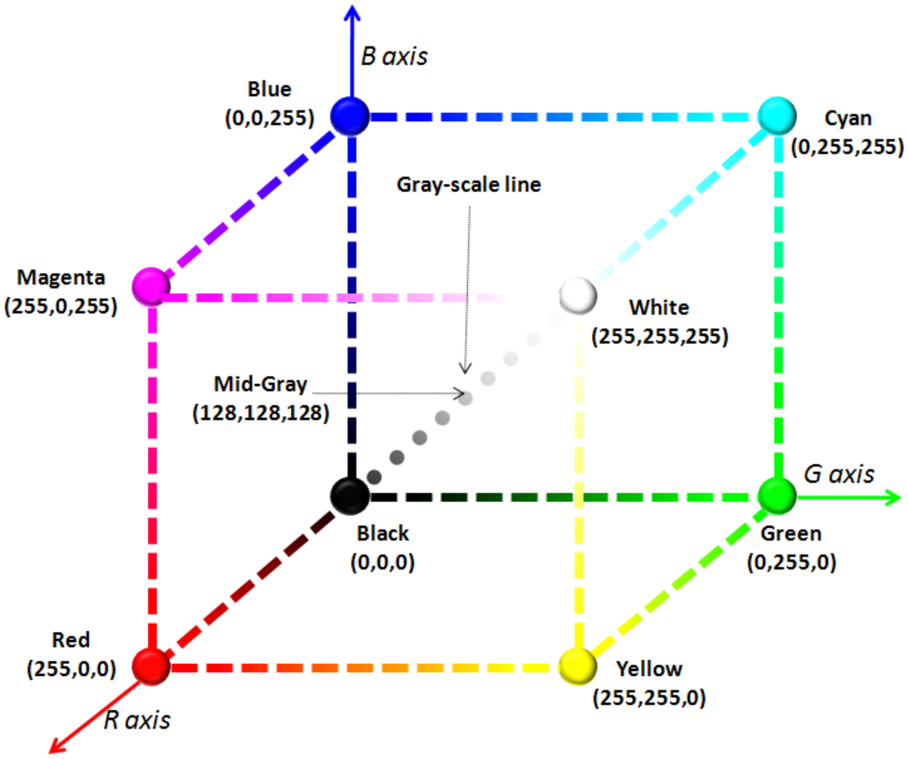

Figure 3.1. 3D orthogonal coordinate system for the RGB color model.

Although suitable for composing digital images, capturing color information from the outside, and displaying colorful images in monitors and television apparatus, the RGB system is not well adequate for describing color in a practical interpretation manner. For instance, humans do not describe object colors in terms of combinations of red, green and blue, but as hues, having washed-out or high-intensity appearances. For this reason, the RGB system was rearranged to create a more perceptually proper color representation, through a cylindrical coordinate system (see Figure 3.2).

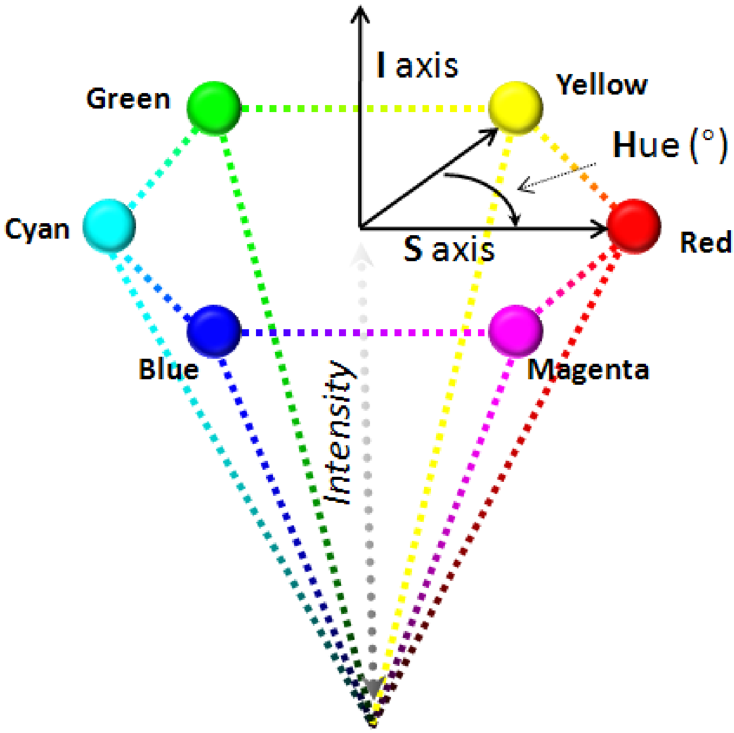

Figure 3.2. Polar coordinate system for the HSI space.

This cylindrical cone-shaped system is known as HSI model (which stands for Hue, Saturation, Intensity). We can readily notice from Figure 3.2 that the RGB gray-intensity diagonal happens to be the middle line denoting the cone height. The red color has the zero degree angle, whereas each other color axis constitutes an angle with relation to the red axis, that is, a single hue is defined by its angle with respect to the red axis. The closer is a point to the outer edge of the cone base (base circle), the purer (more intense) is the color (or hue). Its purity is thus determined by its distance (the radius length) to the gray-scale intensity axis, which is represented by the magnitude of a vector perpendicular to this axis and called saturation. Still at the base circle of the cone, we can see that the closer is the hue coordinate to the axis, the more is it diluted to white, which give us the perception of a pastel like color. In this case, the hue is undefined. On the other hand, if the intensity is closer to the origin (the black point), then the saturation is uncertain.

Typically, the HSI values (*hue*, *sat*, *int*) in terms of the RGB system are calculated by the following equations:

$$hue = \arctan\left(\frac{\sqrt{3}(G - B}{2R - G - B}\right), \tag{3.1}$$

$$sat = 1 - \frac{\min(R, G, B)}{R + G + B}, \tag{3.2}$$

and

$$int = \frac{R + G + B}{3}. \tag{3.3}$$

3.2 Color Fundamentals

Color is a important property in computer vision applications, for instance, image segmentation [Klinker *et al.* , 1989; Geusebroek *et al.* , 2001], object recognition Gevers & Stokman [2004]; van Gemert *et al.* [2006]; van de Sande *et al.* [2010], stereo matching and motion analysis [Klinker *et al.* , 1991]. It is an essential element to readily distinguish and extract objects from scenes by both humans and algorithms. The full potential of color information has not been explored before the community addressed the difficulties imposed by the complex reflectance process underlying image formation. This scenario started to change after the introduction of the dichromatic reflection model by Shafer [1985]. This model explains in simple terms how the physical phenomena caused by the interaction of light in the environment, such as shadows and specularities, impact the formation of RGB-color images in the image sensors. On the basis of such model, several algorithms were developed to construct robust color histograms, holding invariance properties for the reflectance physical effects, to be applied for automated recognition of scenes and objects [Gevers & Stokman, 2004; van Gemert *et al.* , 2006; van de Sande *et al.* , 2010]. Built on the success of those works, we attempt to boost the automatic identification of video events through the use of local spatiotemporal features based on the zero-order color invariants described in [van de Weijer & Schmid, 2006].

This section is thus intended to describe the color theory underlying the design of the photometric color invariants. An overview of the physics of reflection and a detailed description of the mathematical reflectance model are first presented. In this discussion, we also mention about some of the assumptions and limitations of the model. In what follows, we present the two color systems used and their photometric color invariants with basis on the dichromatic reflection model.

3.2.1 Reflection and the Dichromatic Model

The dichromatic reflection model gives a formal representation of the interaction of light with surfaces. It considers inhomogeneous surfaces of dielectric materials such as plastics, paints, ceramics (e.g., porcelain), and paper. Based on the physical reflection properties of these materials, the characteristics of the sensors and the shapes of the objects, this model decouples the spectral radiance into two complementary components, namely one accounting for color highlights (interface reflection) and the other describing color shadings (body reflection).

3.2.1.1 The Physics of Reflection

In real scenes, the color variations seen across object surfaces are to a large extent caused by the effects from the optical reflection process, like specular (or interface) reflection, diffuse (or body) reflection, shading and shadows. These reflection phenomena can be summarized by the occurrence of two concurrent processes: i) the interface reflection, and ii) the body reflection. In the literature, the effects of interface reflection are also referred to as specular reflection, while body reflection as diffuse. In Figure 3.3 illustrates those processes.

The reflection originated from where the surface reaches the air (interface) is termed interface reflection, which is responsible for the white shiny points projected on the surfaces. In contrast, the light reflected off the surface medium produces the diffuse shading reflection, which comes from the colorants inside the materials. The absorption of certain wavelengths from the incoming light by those small particles (the colorants) is what determines the perception of coloration by the observer.

When the surface normal bisects the angle formed by the incident illumination and the direction of light reflected off the interface, an event called the perfect specular reflection occurs (also known as highlight). In this case, the highlight, that is the specular reflection of the light source, appears in its most expressive form. The intensity degree and locale of highlights vary according to the viewing geometry. Thus, not only is specular reflection dependent on lighting geometry but also viewpoints. Highlight produces a confounding characteristic, since it retains nearly the same colour of the illuminant and blends it to the real color of the material, therefore the effect introduced is undesirable. Conversely, diffuse (or body) reflection is known to be generated from the inner scattering of the incoming light by the colorants, which eventually re-emit it to the outside at an arbitrary orientation. This way, the diffuse reflection conveys the shading effects more related to the true coloration of the surface objects.

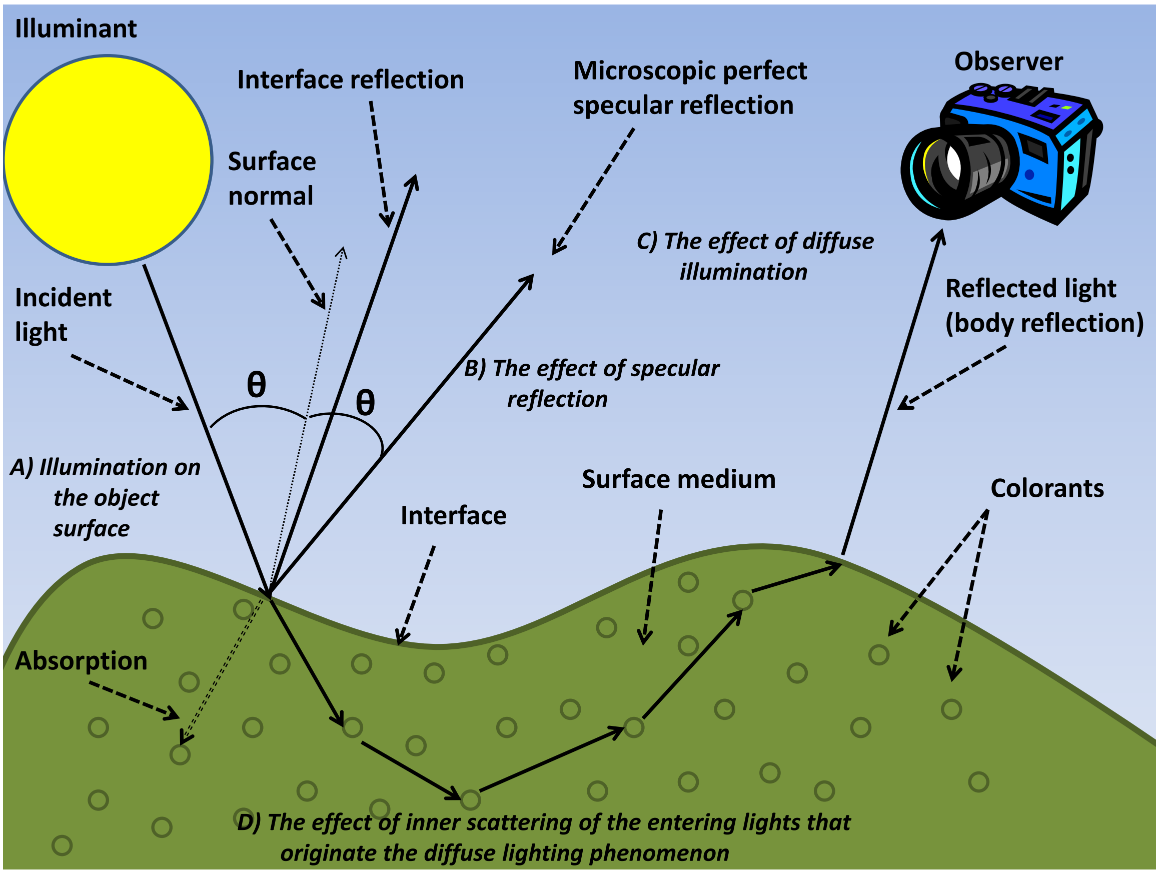
Illuminant
Interface reflection
Microscopic perfect specular reflection
Observer
Surface normal
C) The effect of diffuse illumination
Incident light
Reflected light (body reflection)
θ
θ
B) The effect of specular reflection
A) Illumination on the object surface
Surface medium
Colorants
Interface
Absorption
D) The effect of inner scattering of the entering lights that originate the diffuse lighting phenomenon

Figure 3.3. This illustration depicts the basic dynamics of the reflectance process. An incoming illumination hits the green surface, which originates the various reflectance phenomena, namely body reflection, interface reflection, specular reflection.

Separating the effects of shading from highlight can be very useful. For example, in image segmentation it is generally assumed smooth or uniform intensity variation across the surface, therefore it would be interesting to avoid highlights and work only on the shading image. Another example is depicted in stereo and motion analysis, which count on matching of points from images of different viewpoints, that would benefit from shading intrinsic images.

3.2.1.2 The Dichromatic Model

By considering the two previously examined reflection phenomena, the dichromatic reflection model can be expressed by the linear combination of both events. Given the reflectance geometry depicted in Figure 3.4, in the dichromatic model the total radiance of the reflected light is calculated as

$$L(\lambda, i, e, g) = L_b(\lambda, i, e, g) + L_i(\lambda, i, e, g), \tag{3.4}$$

where i is the angle of incidence, e is the angle of emittance, g is the angle formed by the incidence illumination and the viewing direction, and λ is the wavelength of the reflected light. In this formulation, L_b accounts for the light emitted from the colorants within the surface medium, while L_i denotes the light reflected off the interface.

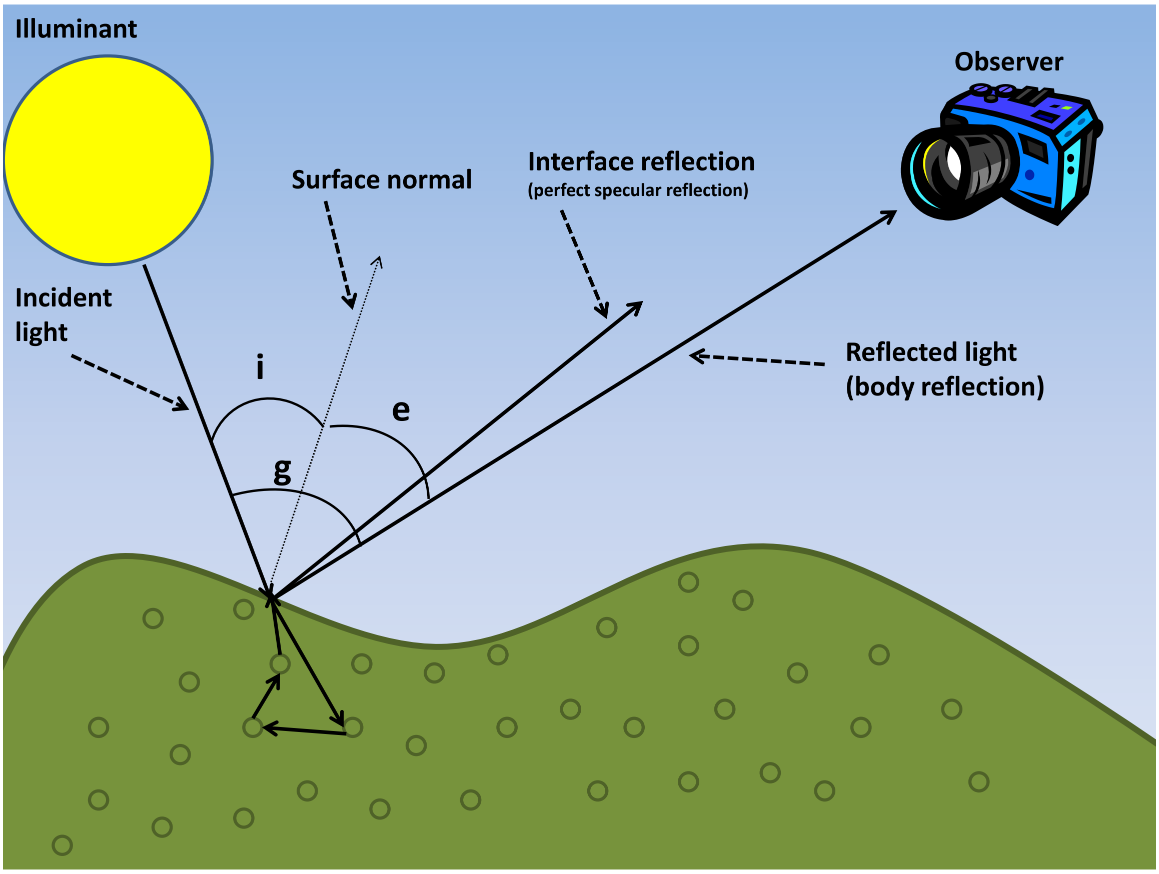

Figure 3.4. This picture details the geometry of lighting in which the dichromatic reflection model find basis.

This model also establishes that each component can be separated in two parts: a geometric scale factor (magnitude) and the relative spectral power distribution (composition); then the equation becomes

$$L(\lambda, i, e, g) = m_b(i, e, g)c_b(\lambda) + m_i(i, e, g)c_i(\lambda). \tag{3.5}$$

Putting that into words, it says that the color of the reflected light has a spectral distribution (c_b or c_i) only depending on the wavelength λ scaled by a magnitude factor

(m_b or m_i) that changes with variations in lighting and viewing directions.

Several assumptions are made regarding the dichromatic model. It limits its scope to opaque, inhomogeneous surfaces, in which the colorants are regularly distributed. Surface reflection is independent of rotation changes about the axis of the surface direction.

Those premises are typical when constructing reflectance models and are said to be plausible approximations to a realistic scenario. However, other two conditions to which the model is constrained are the absence of inter-reflection and the nonexistence of diffuse (ambient) illumination, which are actually very common events in real scenes, having possibly a great portion of influence.

Nevertheless, the model also does not adopt a couple of preconditions, which allows it to possess some flexibility: i) body reflection is not considered isotropic, ii) uniformly spread lighting is not taken into account, iii) the model can be applied to rough, planar and curved surfaces, among others. Only some were mentioned, as our intention is just to provide the reader with a cursory description of the model; we refer the interested reader for more details in [Shafer, 1985; Klinker *et al.* , 1991].

By combining the dichromatic reflectance model with the spectral projection, the process by means of which pixel values are determined from the spectral power distribution of light, a color space defining the pixel colors from the measured amount of reflected light is introduced. In this color space, the final color value results from the summation of the interface reflection vector and the body reflection vector weighed by their corresponding magnitudes (a scalar value).

When the observer is a monochrome camera, the pixel value p is specified by the addition of the amount of incoming light at every wavelength $s(\lambda)$, pondered by the camera's wavelength-based responsivity $\rho(\lambda)$; so $p = \int s(\lambda)\rho(\lambda)d\lambda$. But if a color camera is used, three filters $k_C(\lambda)$ $(C \in (R, G, B))$ is in charge of reacting to three distinct wavelengths (standing for the red (R), green (G) and blue (B) lights), simulating the process of color perception performed in human beings, to estimate their portions of light transmitted. The responsivity function is then replaced by the product of the transmittance function and itself $(f_C(\lambda) = k_C(\lambda)\rho(\lambda)$, so $p = \int s(\lambda)k(\lambda)\rho(\lambda)d\lambda)$. To close the reasoning, each pixel $p(x, y)$ forming an image $I(x, y)$ turns out to be a three-value vector $C_{x,y} = [R_{x,y}, G_{x,y}, B_{x,y}]$, and the pixel function becomes

$$C = \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} \int s(\lambda)k_R(\lambda)\rho(\lambda)d\lambda \\ \int s(\lambda)k_G(\lambda)\rho(\lambda)d\lambda \\ \int s(\lambda)k_B(\lambda)\rho(\lambda)d\lambda \end{bmatrix}. \tag{3.6}$$

By applying the spectral projection model, the reflected light captured by the sensor at a certain position will have the measured color given by

$$C^L_{x,y}(\lambda, i, e, g) = m_b(i, e, g)C^b_{x,y}(\lambda) + m_i(i, e, g)C^i_{x,y}(\lambda). \tag{3.7}$$

In order for the dichromatic reflectance model to take into account diffuse (or ambient) illumination, Shafer only added a diffuse light term, rewriting the Equation 3.7 to

$$C^L_{x,y}(\lambda, i, e, g) = m_b(i, e, g)C^b_{x,y}(\lambda) + m_i(i, e, g)C^i_{x,y}(\lambda) + C^a_{x,y}(\lambda). \tag{3.8}$$

This extended model reveals that the diffuse illumination is independent on geometry, and so it will only be represented by the color of the diffuse light $C^a_{x,y}(\lambda)$. Yet, what this new term does is to shift the parallelogram formed by the two other reflectance terms from the origin of the RGB space.

3.2.2 Photometric Robustness: Color Invariance

This section is concerned with the derivation of the color invariants used in this work. We have only considered zero-order invariants, naturally our discussion in bounded on that, but we refer the reader to [van de Weijer & Schmid, 2006] for an explanation on the first-order invariants. van de Weijer & Schmid have modeled the illuminant as a single local source $i(\lambda)$, and the case of multiple sources was assumed to be an approximation of such as well. This way, the Shafer's equation becomes

$$C(\theta_s, \theta_i, \theta_v) = m_b(\theta_s, \theta_i) \int_w c^C_b(\lambda)e(\lambda)f^C(\lambda)d\lambda + m_i(\theta_s, \theta_i, \theta_v) \int_w c_i(\lambda)e(\lambda)f^C(\lambda)d\lambda, \tag{3.9}$$

such that $C \in (R, G, B)$, $f(\lambda)$ is the sensor sensitivity as a function of the wavelengths from the visible spectrum w, the illumination intensity is given by $e(\lambda)$, m_b is the geometric scaling factor for the surface albedo (which is geometrically dependent of the surface orientation θ_s and the illuminant direction θ_i), m_i is the the geometric scaling factor for the interface reflection (which has the same geometry dependencies of the body reflection plus the viewing direction θ_v), c_b and c_i account for the colors of body reflection and interface reflection, respectively, both depending only on the wavelength λ.

By assuming a white illuminant, meaning that the energy distribution over the visible spectrum is the same at every wavelength, i.e. $e_R = e_G = e_B = e$, $e(\lambda) = e$ is constant; and also $f(\lambda_R) = f(\lambda_G) = f(\lambda_B) = f$. Furthermore, by making two more assumptions, namely i) the differences of refractive indices among all wavelengths are negligible, so nearly constant, and ii) consequently the specular reflection color can be approximated to the illuminant color, that is, the neutral interface reflection (NIR) is considered, $c_i(\lambda) = c_i$ also becomes constant. As a result, the Equation 3.9 above can be written as

$$C(\theta_s, \theta_i, \theta_v) = m_b(\theta_s, \theta_i)e \int_w c_b^C(\lambda) f^C(\lambda) d\lambda + m_i(\theta_s, \theta_i, \theta_v) c_i e f. \tag{3.10}$$

Supposing matte surfaces (that is, no interface reflection occurs, so $m_i(\theta_s, \theta_i, \theta_v) = 0$) in the scene, under the same conditions described, and that the sensors approximates $f(\lambda_C)$ by the delta function (i.e., $f^C(\lambda) = \delta(\lambda - \lambda_C)$), we have that the equation is reduced to

$$C(\theta_s, \theta_i, \theta_v) = m_b(\theta_s, \theta_i) e c_b^C(\lambda). \tag{3.11}$$

Note that the equation is independently applied to each channel, as $C = (R, G, B)$. Clearly, $R = m_b(\theta_s, \theta_i) e c_b^R(\lambda)$, $G = m_b(\theta_s, \theta_i) e c_b^G(\lambda)$ and $B = m_b(\theta_s, \theta_i) e c_b^B(\lambda)$, and it is possible to compute RGB values invariant to the intensity of the light source, the lighting geometry and viewpoints by normalizing each channel by the summation of the three channel values. Then, the equations are

$$R = \frac{m_b e c_b^R}{m_b e (c_b^R + c_b^G + c_b^B)} = \frac{c_b^R}{c_b^R + c_b^G + c_b^B}, \tag{3.12}$$

$$G = \frac{m_b e c_b^G}{m_b e (c_b^R + c_b^G + c_b^B)} = \frac{c_b^G}{c_b^R + c_b^G + c_b^B}, \tag{3.13}$$

and

$$B = \frac{m_b e c_b^B}{m_b e (c_b^R + c_b^G + c_b^B)} = \frac{c_b^B}{c_b^R + c_b^G + c_b^B}. \tag{3.14}$$

Now, if we still limit the model to a white illuminant but under the occurrence of specular reflection (so $m_i \neq 0$), the opponent colors [van de Weijer & Schmid, 2006] can be demonstrated invariant with regard to specularities, even though variant to geometry and the incident light intensity, once

$$O_1 = \frac{1}{\sqrt{2}}(R - G) = \frac{1}{\sqrt{2}}(m_b e (c_b^R - c_b^G) + m_i e - m_i e), \tag{3.15}$$

$$O_2 = \frac{1}{\sqrt{6}}(R + G - 2B) = \frac{1}{\sqrt{6}}(m_b e(c_b^R + c_b^G - 2c_b^B) + 2m_i e - 2m_i e). \tag{3.16}$$

To simultaneously acquire invariance with respect to lighting geometry, viewpoints and specularities, the hue formula is deduced from the color opponents equations as follows,

$$hue = \arctan\left(\frac{O_1}{O_2}\right) = \arctan\left(\frac{\sqrt{3}(c_b^R - c_b^G)}{c_b^R + c_b^G - 2c_b^B}\right). \tag{3.17}$$

It is known from [Gevers & Stokman, 2004] that the color invariants own instabilities inherited from their transformations. Since those instabilities are not wanted to affect the discriminative power and robustness of color histograms, van de Weijer & Schmid [2006] apply the error analysis studied in [Gevers & Stokman, 2004] as a weight to the construction of the color descriptors. In particular, the hue values are unstable at the gray axis. From the hue error analysis (see Equation 3.18), the certainty of the hue is inversely proportional to the saturation, i.e. the lower the saturation, the more uncertain is the hue value. Robustness to this can be accomplished by weighing each hue sample by its respective saturation (given by the Equation 3.19), as proposed in [van de Weijer & Schmid, 2006].

$$(\partial hue)^2 = \left(\frac{\partial hue}{\partial O_1}\partial O_1\right)^2 + \left(\frac{\partial hue}{\partial O_2}\partial O_2\right)^2 = \frac{1}{O_1^2 + O_2^2} = \frac{1}{sat^2} \tag{3.18}$$

$$sat = \sqrt{O_1^2 + O_2^2} = \sqrt{\frac{2((c_b^R)^2 + (c_b^G)^2 + (c_b^B)^2 - c_b^R c_b^G - c_b^R c_b^B - c_b^G c_b^B)}{3}} \tag{3.19}$$

In a nutshell, this section explained how the color descriptions can acquire invariance properties under assumptions by the dichromatic reflection model. Table 3.2 summarizes this discussion.

3.3 The Proposed Methods

This work extends the Laptev's space-time interest point detector to incorporate color information by the early fusion approach, in which two modifications were proposed. First, instead of the intensity gray-scale channel, the three *rgb* color channels were input to the interest point operator; this version was named the ColorSTIP. Second,

Table 3.1. A summary of the invariant properties for the discussed color models. *rgb* is the *normalized*-RGB color system, and *hue* is the H component of the HSI system derived from the opponent color model. The symbol + means to hold robustness to a specific event, whereas the symbol − means the lack of robustness.

color model	***viewing direction***	***surface orientation***	***specular reflection***	***illumination direction***	***illumination intensity***
rgb	+	+	−	+	+
hue	+	+	+	+	+

the support local space-time regions were also described in terms of the *hue* values, such that the color histograms were concatenated to the feature histograms; HueSTIP was the name given to this version. An intuitive yet interest adaptation was to join both versions in a single one, which we have called the Hue-ColorSTIP. As follows, we give a discussion on the two schemes used to incorporate color information to STIP, which originated the family of color-based STIPs.

3.3.1 ColorSTIP (Color-based Spatiotemporal Interest Points)

In this version, we try to detect space-time interest points also robust to photometric variations. To this end, for each frame, we replace the gray-scale channel by the three channels from the *rgb* system. So, the Gaussian derivatives are applied to each color channel at all directions (the $x, y,$ and t dimensions). After that, the final color-based Gaussian derivative at each dimension is obtained by the summation of its Gaussian derivatives from each color channel, which becomes

$$L_x^{rgb} = L_x^r + L_x^g + L_x^b, \tag{3.20}$$

$$L_y^{rgb} = L_y^r + L_y^g + L_y^b, \tag{3.21}$$

$$L_t^{rgb} = L_t^r + L_t^g + L_t^b. \tag{3.22}$$

The *rbg*-based second moment matrix μ^{rgb} from the space-time extension of the Harris-Laplace corner detector will then be given by

$$\mu^{rgb} = g(.;\sigma_i^2,\tau_i^2) * \begin{pmatrix} (L_x^{rgb})^2 & L_x^{rgb}L_y^{rgb} & L_x^{rgb}L_t^{rgb} \\ L_x^{rgb}L_y^{rgb} & (L_y^{rgb})^2 & L_y^{rgb}L_t^{rgb} \\ L_x^{rgb}L_t^{rgb} & L_y^{rgb}L_t^{rgb} & (L_t^{rgb})^2 \end{pmatrix}, \tag{3.23}$$

while the rest of the method continues the same. By doing this, the original STIP starts to hold properties that it lacked before. The properties to which we refer stem from the photometric invariant properties held by the *rbg* system, namely lighting geometry, viewpoints and illumination intensity. With those additional characteristics, the ColorSTIP is expected to become an improved version of the STIP.

3.3.2 HueSTIP (Hue histograms added to STIP descriptors)

This version arose from the combination of histograms of hue describing the space-time interest points with the other two feature descriptors accounting for motion (HoF) and appearance (HoG). So, basically, each space-time interest point is also represented in terms of the hue values quantized at the local spatiotemporal region of the point. The range of values for the hue is usually measured as angles in radians $(0 - 2\pi)$ and we divide this range into 36 bins to follow something similar to [van de Weijer & Schmid, 2006]. This lets each bin accounts for one range of hue values.

To construct the hue histogram, we calculate the *bin* number to which the *hue* value (of a pixel in the spatiotemporal volume) belongs with the formula $bin = hue * 36/2\pi$. Then, the saturation value at corresponding pixel is accumulated at the position *bin* to which the *hue* value was assigned. Before accumulating the amount of saturation in the histogram bin, the saturation is weighed by a corresponding value in a weighing Gaussian mask. This means that depending on the position of the pixel, the saturation will have a different weight for participation. For the pixels centered at the spatiotemporal volume, the saturation will have total participation when quantizing the hue histogram. The size and values forming the spatiotemporal Gaussian mask will vary according to the spatial and temporal scales of the interest point.

By using the hue histograms to describe the spatiotemporal features, the color information should aggregate robustness to the extracted patterns in terms of illumination intensity, lighting geometry, viewpoints and specular reflection.

3.3.3 Hue-ColorSTIP

This last version is the joint of the two color-based STIPs first described. What motivated us to proposed such combination was supported by the hypothesis that the color description on intensity-based interest points may not be appropriate as the points were not detected by using color information. We think that maybe it can be more powerful to describe the interest points in terms of color if they are detected by using the ColorSTIP.

3.3.4 Summary

Table 3.2 summarizes the invariance properties acquired by each color-based STIP extension. We show that HueSTIP has robustness to photometric variations to which the *hue* channel is unaffected. However, the invariance properties of the HueSTIP only apply to the representation of the interest points. On the contrary, the ColorSTIP is invariant to all phenomena, but only at the level of the interest points. The Hue-ColorSTIP is robust to all of the photometric variations in both concepts of local features (interest points and descriptions), but specular reflection at the level of interest points.

Table 3.2. A summary of the invariant properties incorporated by the family of extensions of STIP. The first column lists all STIP's versions. Each remaining column accounts for one specific photometric effect. The symbol + means to hold robustness to a specific event, whereas the symbol − means the lack of robustness. The left side of / refers to the space-time interest points, while the right side of / refers to the description of such points.

	photometric phenomena				
proposed extension	***viewing direction***	***surface orientation***	***specular reflection***	***illumination direction***	***illumination intensity***
HueSTIP	−/+	−/+	−/+	−/+	−/+
ColorSTIP	+/−	+/−	−/−	+/−	+/−
Hue-ColorSTIP	+/+	+/+	−/+	+/+	+/+

Chapter 4

Experiments and Evaluations

In previous chapters, we provided the theoretical basis for the proposed methods. We first justified our choice regarding the use of local space-time interest points to represent motion events. Secondly, we gave details about how this method can localize space-time features, and which techniques can be used to portray the detected features in terms of shape and motion appearances. At last, the color theory dictating the invariance properties underneath the target color systems was explained and the family of color-based STIPs was briefly presented.

In this chapter, we explain the experimental framework utilized to the qualitative evaluation of the revised STIPs (HueSTIP, ColorSTIP, and Hue-ColorSTIP) and discuss the results obtained from those experiments. As our objective is to process image sequences showing realistic and complex scenarios, the dataset selection comprised a movie dataset from the literature (namely the Hollywood2 [Marszałek *et al.* , 2009] dataset) and two other specialized datasets constructed during the development of this dissertation. All datasets will be detailed in Section 4.1.

The experimental framework is unfolded in Section 4.2, which was split into three main subsections. The first subsection fully describes the *bag-of-features* approach that, using the low-level features as the primitive data source, was applied to provide a more meaningful representation to the videos. Action classification, the target application context of this work, is based on a supervised learning approach, which will be next set out. To close the section, an outline of the experimental protocol is presented, listing the steps taken to accomplish the battery of experiments.

Lastly, we show the experimental results and discuss them in separate views, lined up in study cases. Each study case is in turn related to the methods' evaluation on one of the target datasets.

4.1 Datasets

Three datasets of realistic scene videos were used in this work (see Table 4.1). The Hollywood dataset first appeared in [Laptev *et al.* , 2008a], where the concept of spatial pyramids was generalized to the spatiotemporal domain. That work defined a multi-channel classification approach, where each combination of one feature type (HoG or HoF) with a spatiotemporal grid was considered a channel. They tested a diverse set of those combinations and were able to improve classification rates for individual human action classes. One year later, the Hollywood2 dataset was presented in [Marszałek *et al.* , 2009], positting a joint framework for action and scene recognition.

The Violence dataset was first introduced in [de Souza *et al.* , 2010], with a violence detector based on the concept of spatiotemporal visual codebooks combined with linear Support Vector Machines. In [Valle *et al.* , 2011], we proposed a simple and robust automated filter of unwanted content (such as violence and pornography), also introducing the Pornography dataset.

It is important to note that for the specialized cases, to the best of our knowledge, there is no shared dataset in the scientific community. A difficulty that obviously arises is how to label the videos, a task which is not only laborious, but also extremely subjective (what is pornography? what is violent content?). To handle this issue, we have collected the positive examples in each dataset from online video-sharing networks which specialize in the kind of content of interest (pornography or fights)

For the negative classes, we gathered videos from general-public social networks. Each video was then inspected by a team of researchers to ensure that no pornographic or violent content was indeed present: again, this introduces a degree of subjectivity in the dataset, but this was unavoidable since even in general-purpose social networks this kind of content sometimes creeps in.

4.1.1 The Datasets of Hollywood movies

We wanted to evaluate the performance of the descriptors for human action recognition in natural scenarios. Therefore, the Hollywood2 dataset [Marszałek *et al.* , 2009] was a natural choice. It is composed by 12 actions: answering phone, driving car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, standing up (see Figure 4.1). Videos were collected from a set of 69 different Hollywood movies, where 33 were used to generate the training set and 36 the test set. Action video clips were divided in three separate subsets, namely an automatic (noisy) training set, a (clean) training set and the test set. The clean training dataset

Table 4.1. Summarized description of the datasets.

Dataset	Description
Hollywood2	This is an extension of the previously proposed Hollywood dataset [Laptev *et al.* , 2008a], in which the number of clean training videos increased from 231 to 823 and the test set changed from 217 to 884 samples. This growth of the dataset was due to the addition of four new action classes (summing up to 12 classes), as well as to the increase in samples for the existing classes.
Violence	The video clips of violent scenes were all retrieved from social networks specialized in fights. It features a diverse set of situations as depicted in Figure 4.3. In total, there are 200 videos of violence and 200 "negative" videos without violence.
Pornography	This dataset contains 400 pornographic videos collected from specialized social network and 400 "negative" videos collected from general public social networks. From the negative videos, 200 are especially challenging, for containing a lot of exposed skin.

was formed by clips manually cut off of the videos, while for the noisy training set an automatic procedure was used for selecting the action clips. We only used the clean training set containing 823 samples and the test set containing 884 samples (see Table 4.2 for more details on the number of videos each action class).

4.1.2 Pornography Database

The Pornography dataset attempts to portray the content diversity found on sharing social networks. For the pornographic class, videos were retrieved from websites specialized on the genre. The content varies in terms of the ethnicities (asians, blacks, whites, multi-ethnic) of the people appearing in the scenes, as well as with regard to the sexual orientation and sexual practices. The authors sampled 400 videos of the porn class, but for our experiments we have randomly selected 200 of them.

For the nonpornographic class, the video content comprised many different subjects: documentaries, educational videos, car races, TV programs, cartoons, music concerts, sport matches, dancing, interviews, daily news, among many others. Of the sampled 400 nonporn videos, 200 exhibit confounding characteristics related to the pornographic content, for example, women wearing bikinis at the beach and undressed babies being bathed. In those cases the exposure of skin imposes a challenge to the

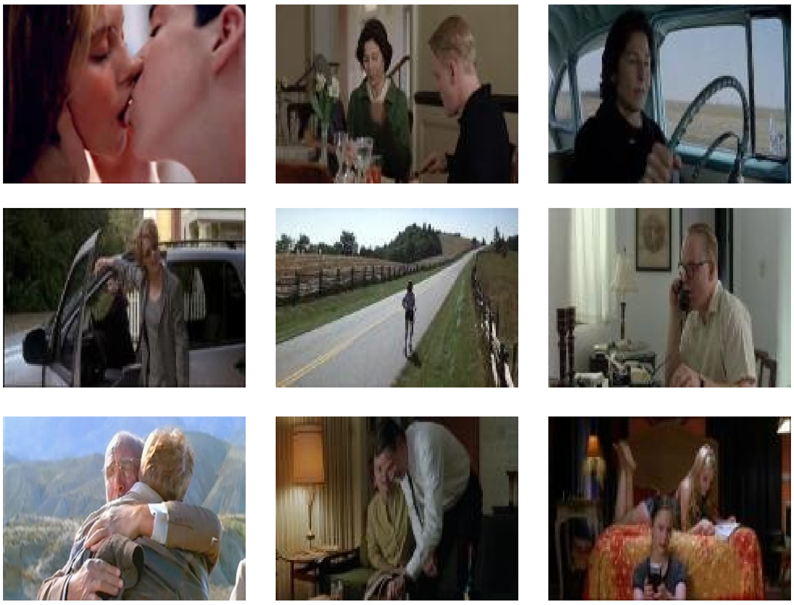

Figure 4.1. Illustration of scenes from the human action dataset.

system. On the other hand, the other half of the nonporn set (more 200 videos) display contents totally visually unrelated to pornographic scenes. We also only selected 200 videos from the entire pornography dataset for our experiments. An illustration of the dataset is depicted in Figure 4.2.

4.1.3 Violence Database

The Violence dataset was one of the contributions of this work. We collected video clips from social networks specialized in fights (Figure 4.3). Both violence and nonviolence samples try to be diverse and representative. A compilation of daily life situations in schools, ghettos, entrance spaces of night clubs, matches from several sport variations (*e.g.*, soccer, hockey), traffic, involving both spontaneous fights and professional wrestling sports build the violence dataset. Scenarios are depicted by aggressive behaviors involving any number of people, in indoor and outdoor environments, with or without presence of moving objects in the background (*e.g.*, cars). The nonviolence dataset was created by copying from the negative class of the Pornography dataset videos that do not exhibit violent content. In total, there are 400 videos, 200 from each category.

Table 4.2. This a description on the number of videos in each class of the Hollywood datasets.

Dataset:	**HOHA**		**Hollywood2**	
Action	*test*	*train*	*test*	*train*
AnswerPhone	23	22	64	66
DriveCar	-	-	102	85
Eat	-	-	33	40
FightPerson	-	-	70	54
GetOutCar	13	13	57	51
HandShake	19	20	45	32
HugPerson	22	22	66	64
Kiss	51	49	103	114
Run	-	-	141	135
SitDown	30	47	108	104
SitUp	10	11	37	24
StandUp	49	47	146	132
Average	217	231	884	823

4.2 Experimental Setup

In our experiments, we investigated the power of the spatiotemporal local features containing color information for action recognition. This section describes the methods and metrics used in the experimental setup to perform the whole set of experiments and evaluations. In part, the experimental framework was inherited from the literature, a recognition protocol that has been widely deemed as a powerful tool to tackle the action classification task. It complies the use of the standard bag-of-features paradigm for a less primitive representation of videos and the Support Vector Machines (SVM) for the classification task. Regarding this latter, our protocol differs by the use of the one-against-one approach and the linear kernel (to make the classifier the simplest possible and highlight only the power of the features). As follows, the approach for the mid-level representation of videos is explained. Then, we present the approach employed for classification and some discussion about its basic concepts.

4.2.1 Bag-of-Features Video Representation

When spatiotemporal local features are extracted, they only provide a very local and unstructured representation of the video clips. One way to give a more meaningful representation is to use the *bag-of-features* (BoF) approach, which has been successfully

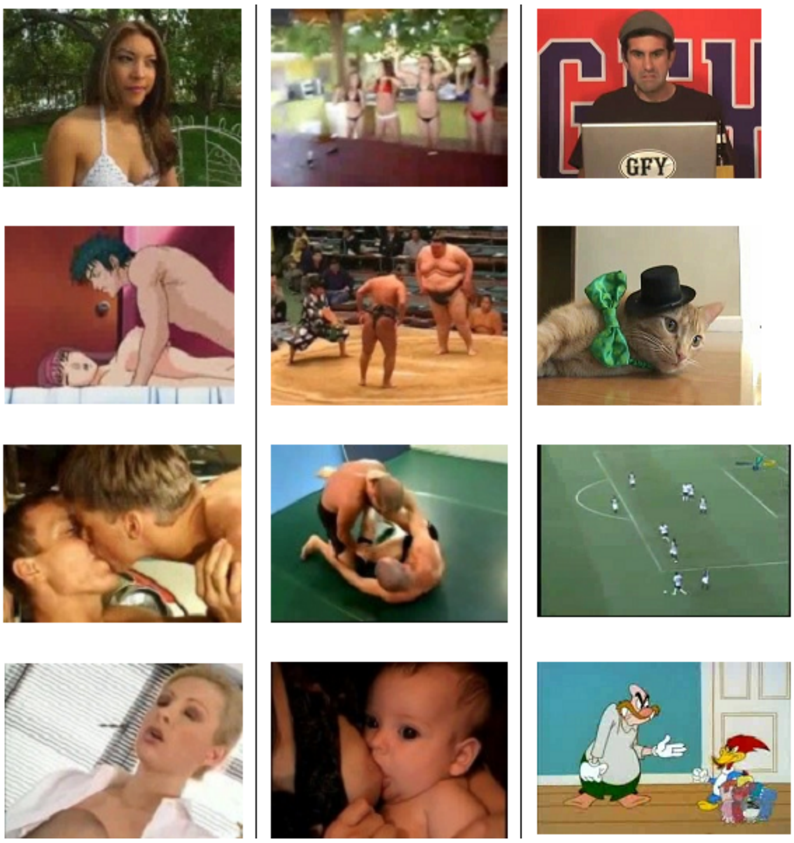

Figure 4.2. Pornography dataset illustration. At the left, only videos of pornographic content. At the center, samples of difficult cases of nonpornographic content. At the right, we show the easy cases of nonpornographic content.

applied to many applications of video analysis [van de Sande *et al.* , 2010; Laptev *et al.* , 2008a]. Bag-of-Features is a method for representing images or videos by histograms of the occurrence rates of patterns called visual words. This is an idea borrowed from the concept of Bag-of-Words (BoW) from the field of textual information retrieval.

Using BoF requires the construction of a vocabulary of features (or visual voabulary). Although this is commonly accomplished by using k-means, it is well known that for very high-dimensional spaces, simple clustering algorithms perform badly, and thus a reasonable and efficient choice is just to select a random sample to form the visual vocabulary: this saves computational time and achieves comparable results [Viitaniemi & Laaksonen, 2008]. The selection or clustering of features to produce the visual vocabulary is performed on the features of the set of training videos.

The vocabulary size K was set to 4000, since this number has empirically demonstrated good results and is consistent with the literature [Laptev *et al.* , 2008a; Marszałek *et al.* , 2009]. This means that by using the k-means algorithm, 4000 will be the number of cluster to be formed having the training features as input, while by the

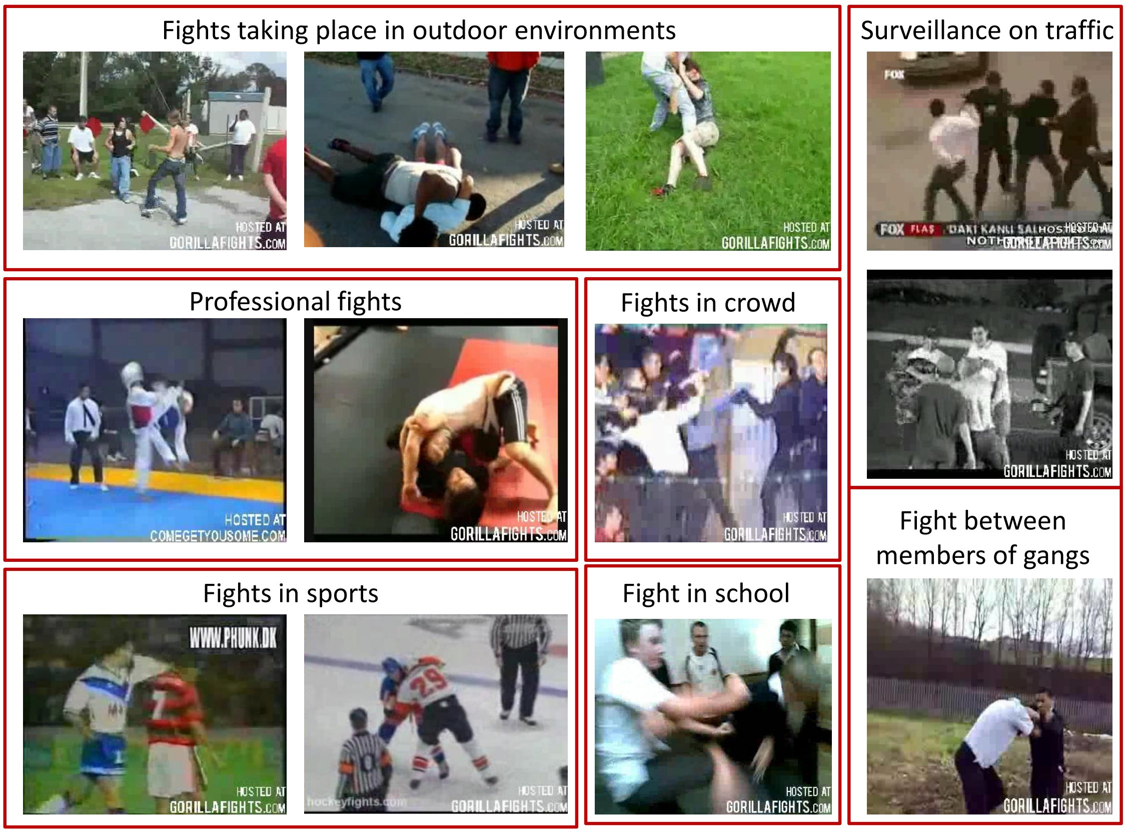
Fights taking place in outdoor environments
Surveillance on traffic
Professional fights
Fights in crowd
Fight between members of gangs
Fights in sports
Fight in school
HOSTED AT GORILLAFIGHTS.COM
HOSTED AT COMEGETYOUSOME.COM
WWW.PHUNK.DK
hockeyfights.com
FOX

Figure 4.3. Violence dataset depicted by several scenarios of fights.

randomly selection mode, 4000 arbitrarily collected from the training features will be used as the visual words (analogous to the centroids of k-means clusters). Then, spatiotemporal local features contained in a (training or test) video clip is assigned to the closest visual word of the visual vocabulary (each assignment is a count in the k^{th} position of the video's visual histogram), using the Euclidean distance. Those counts will form a histogram representation for the videos based on the occurrence rate of the visual words.

4.2.2 Classification

In the literature, the problem of human action recognition has been commonly addressed by machine learning techniques, which has involved the use of supervised learning methods such as Nearest Neighbor Classifier (NNC) [Oikonomopoulos *et al.* , 2006; Wong & Cipolla, 2007; Rapantzikos *et al.* , 2009; Dollár *et al.* , 2005], Support Vector Machines (SVM) [Schuldt *et al.* , 2004; Ke *et al.* , 2007; Laptev *et al.* , 2008a; Willems *et al.* , 2008], and Linear Programming Boosting (LPBoost) [Jiang *et al.* , 2006], as well as unsupervised learning with the Probabilistic Latent Semantic Analysis (pLSA) method in [Niebles *et al.* , 2008]. Supervised learning has apparently provided the best accuracies [Rapantzikos *et al.* , 2009] and therefore we will follow this direction. In particular, to classify the videos, we have used SVM with the libSVM implementation by Chang & Lin [2001].

The multi-classification problem can be addressed by distinct ways yet based on binary classifiers, namely the *one-vs-one*, the *one-vs-all* and the *Directed Acyclic Graph multiclass* approaches (we refer the interested reader to see more details on those approaches in [Hsu & Lin, 2002]). After the work in [Hsu & Lin, 2002], the libSVM's authors chose to implement the multi-classification task with the one-to-one approach for being less time-consuming. In such approach, $n(n-1)/2$ binary classifiers (or binary prediction models) are created (where n is the number of classes) and, after a new element receives labels from all binary models, a majority voting scheme is applied in order to assign the final class label.

Since our aim was to highlight the performance power of the feature types generated by the different extensions, we have decided to simplify the classifier by using the linear kernels for the first battery of experiments, with all class-specific and the Hollywood2 datasets. On the class-specific problem we have performed experiments with a 5-fold cross-validation approach, while for the multi-class problem we used a test-train scheme.

SVM stands for Support Vector Machines and was developed to handle classification and regression analysis problems. In this concept, $C-1$ hyperplanes based on labeled training point samples of different C categories are created for the purpose of separating the target classes the best way possible (the margins between the hyperplanes are maximized). Once the hyperplanes are found using a training set, the prediction model is ready. The characteristics of the hyperplanes will be determined by the SVM kernel, which is, in essence, the key for the effectiveness of the classification model, thus for the prediction stage success. Kernels can be linear or non-linear, and while the linear kernel is the simplest, there are many non-linear kernels that can be used, such as polynomial, sigmoid, RBF (Radial Basis Function), and χ^2 kernels. For action recognition using STIP, the most commonly used is the χ^2 kernel [Laptev *et al.* , 2008a; Marszałek *et al.* , 2009].

4.2.3 Evaluation Criteria

According to Schmid *et al.* [1998], the assessment of interest point detectors requires the analysis of two basic properties: *repeatability* (the fact that the same points are found in a object or scene, even when strong transformations are applied) and *distinctiveness* (points referring to different parts of an object or scene are given easily distinguishable descriptors), which are measured quantitatively. Local descriptors performance has also been assessed through the analysis of Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) graphs in [Mikolajczyk & Schmid, 2005]. Other means of evaluation consist in reporting either the Accuracy or Average Precision (AP) measurements, such as in [Laptev *et al.* , 2008a; Marszałek *et al.* , 2009; Rapantzikos *et al.* , 2009].

In [Wang *et al.* , 2009], accuracies and average precisions were presented in order to assess the performance for different combinations of state-of-the-art detectors and descriptors evaluated on several datasets (including artificial and realistic scenarios). Many works in video classification have followed the evaluation framework by Laptev *et al.* [2008a], such as [Marszałek *et al.* , 2009; Rapantzikos *et al.* , 2009; Wang *et al.* , 2009]. We follow their example in the general experimental setup, though we prefer to employ a straightforward linear kernel SVM, that emphasize the difference between the quality of the feature spaces (a sophisticated nonlinear kernel is able to extract more information from an "unruly" feature space). It is also noteworthy that currently, linear machines are also more scalable to very large instances than those which employ nonlinear kernels, and this is important in the social network context.

Another very common tool used to evaluate classification performance is the confusion matrix. Given a binary classification problem, the matrix confusion is defined as in Table

Table 4.3. Template of a confusion matrix.

	Prediction (how it was classified)	
Class	*Label1*	*Label2*
Label1	True Positive	False Negative
Label2	False Positive	True Negative

Accuracy indicates how close the system can score the expected values.

$$Acc = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Negative} + \text{True Negative} + \text{False Positive}} \tag{4.1}$$

Recall represents the amount that has been successfully classified, that is, how many items where correctly classified as the positive class.

$$R = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{4.2}$$

Precision represents the fraction of real positive items that were correctly classified among all items classified as positive.

$$P = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{4.3}$$

False Positive Rate is the portion of items that were incorrectly classified as the positive class given all items of the negative class.

$$FPR = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \tag{4.4}$$

4.2.4 The Experimental Protocols

The experiments of this work were performed by means of two protocols. One is closely related to the standard experimental framework of the literature for the multiclass problem of human action recognition, and was employed so the results obtained could be — in theory — compared to those of the literature. The other is a more statistically sound 5-fold cross-validation protocol, which we tried to make as compatible as possible with the first. In this section, we give all steps needed to reproduce those experiment protocols.

4.2.4.1 Experimental Protocol for the Multiclass Problem

This first protocol was used to evaluate the feature algorithm on the multiclass problem of human action recognition (with the Hollywood2 dataset) and its detailed steps are described as follows:

1. Extract the low-level features of all videos in the dataset (e.g., STIP, HueSTIP). Keep in mind that the variation in the low-level feature is the main factor being evaluated in this protocol.

2. For each low-level feature being evaluated:

   a) Construct a visual codebook by using its corresponding training feature set (see Section 4.2.1 to know how to perform this step);

b) Compute the histograms (bags) of visual words for each video clip. This means that every video will have a histogram representation by visual word occurrences;

c) Organize the histograms by class (in the test and the training sets). Keep in mind that, at this point, each histogram represents a video in terms of visual words;

d) Learn a binary classifier class for each pair of classes (this is the one-against-one approach) by using the training set of histograms of visual words. The linear kernel is used to learn the SVM-based classification model;

e) Classify each sample of the test set of visual histograms by each one of the binary classifiers. Apply the majority vote scheme to determine the final class label of the test sample;

f) Check the label of test set against the ground truth and build the test metric (Precision-Recall, confusion matrix, etc.).

While we have tried to be as close as possible to the experimental protocol presented in [Laptev *et al.* , 2008b], which is also based on BoF, ours differs in some implementation details: the kernel employed in [Laptev *et al.* , 2008b] is the χ^2 kernel and multiclassification approach is the one-against-all.

4.2.4.2 Experimental Protocol for the Specialized Cases

For the binary classification problems, Violence and Pornography, we used a 5-fold classification scheme having the first protocol as basis for evaluating each fold. It is described as follows.

1. Take one dataset, then

2. Separate the dataset (approximatelly) equally into 5 subsets, such that the number of videos in each subset is balanced by the two class labels (for example, in the Pornography case, there will be nearly the same quantity of pornographic videos and nonpornographic videos);

3. For each subset,

   a) Consider the current subset as a test set and the other subsets together as a training set;

   b) Apply the first protocol described (see Section 4.2.4.1).

4.3 Results and Discussions

In this section, we present the experimental results on the evaluation of the color-based STIPs against the original, intensity-based version. We also emphasize the experimental procedures and present the parameter values used for the classification framework. The analysis of results is contextualized under different application contexts where one metric can be more important than others. By considering this, we state which algorithm can be more appropriate depending on the scenario's requirements.

4.3.1 Human Action Classification for Video Annotation and Retrieval

The first objective was to verify if the proposed extensions of the STIP could achieve any gain in performance. In this experiment (see Section 4.2.4 for details on the setup), the multiclassification problem is considered mutually exclusive, but keep in ming that in the Hollywood2 dataset a video can contain more than one action (for example, it is natural to find scenes in which kissing also involves hugging, see Figure 4.4). We considered the C-SVC version of the SVM, with the C values chosen by cross-validation on the average accuracy per class from 0.001 to 1000. For those preliminary results, we used only the accuracy and confusion matrices as metrics.

Figure 4.4. Two samples of scenes where people are kissing and hugging at the same time.

The experimental results shows that indeed there is gain in performance when using the color-based feature algorithms, such that ColorSTIP held the best overall performance (see Table 4.4). Looking individually at each class, it is possible to note that for a subset of the classes, all color-based algorithms were superior to the luminaceonly based one (STIP), namely *DriveCar*, *HandShake*, *Kiss*, *SitDown* and *StandUp*. For the *HandShake* and *Kiss* classes, the benefit in the use of color information is explained by the scenes of close-up of hands and faces that in most of the related videos (see Figures 4.5 for an illustration of this fact). In those cases, skin color corresponds to the same significant area of the scenario, which probably helps to better define the two human actions. However, it is important to keep in mind that motion features and shape information remain important to define the human parts performing the actions. Therefore, color information is expected to enhance, not to replace any of the other features.

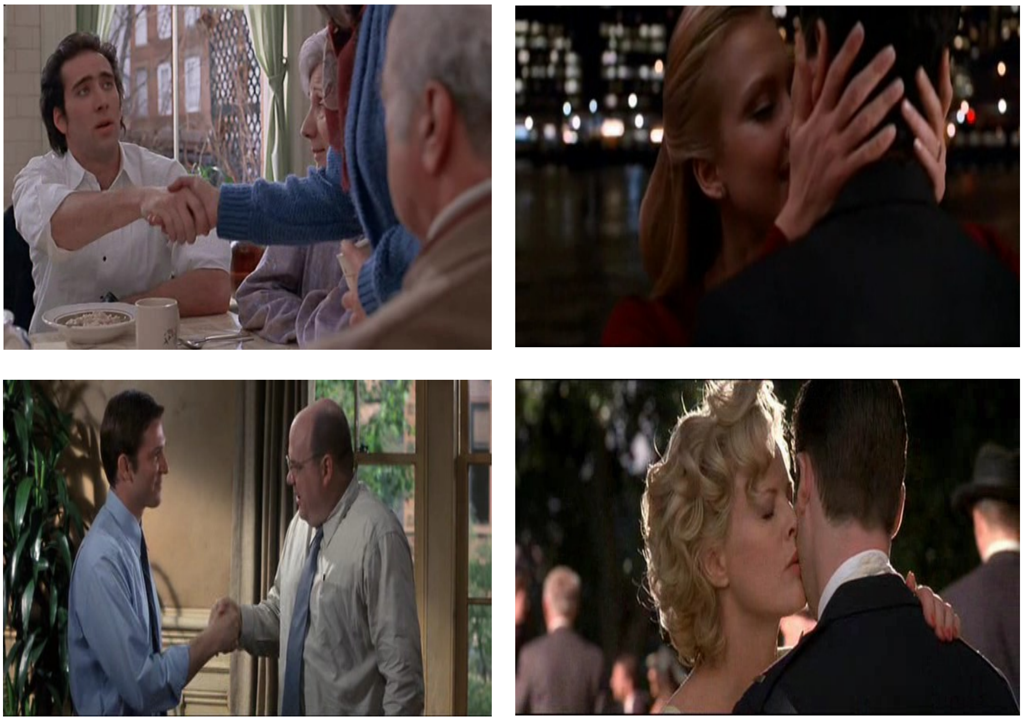

Figure 4.5. Scenes of the *Kiss* and *HandShake* actions in which the objects of interest are focused by the camera.

It is interesting to observe that half of the mistakes made by the ColorSTIP of videos correctly hit by STIP were grayscale clips, as can be seen in Figure 4.6), and in one specific video the skin color is not shown at all (see its illustration in Figure 4.6(G)). As expected the skin color information taken from face close-ups in *Kiss* scenes were very helpful to increase the number of hits by the ColorSTIP (see Figure 4.7).

In the *HandShake* class, we noted that the actions generally happen with people standing, and also can usually involve more than two people. In this sense, color infor-

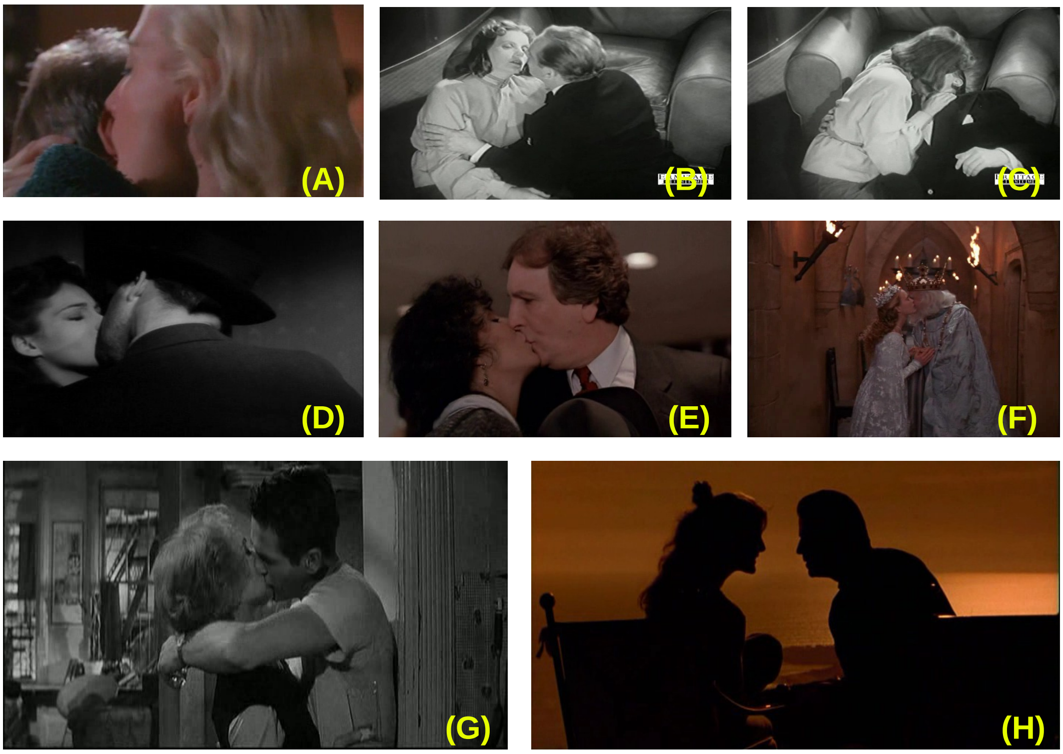

Figure 4.6. Illustration of the set of *Kiss* action videos that were correctly classified by STIP, but not by ColorSTIP. Figures (B), (C), (D) and (G) are in gray scale, while the others are colorful. Figure (G) only depicts the sillhuoettes of the actors, so no skin color is shown.

mation can be viewed as an important element to describe the objects that generally appear in a *HandShake* scene. In addition, the backgrounds from *HandShake* scenes vary a lot, especially in what concerns the color information describing the action context. As a result, the color information embedded in the motion pattern detector impacted by a performance increase of 6.66% (by contrasting ColorSTIP with STIP), while the color description gave only a slight improvement.

Among other characteristics, *DriveCar* scenes also involve the close-up of faces from the people inside the car (see Figure 4.8). Naturally, color can be useful to improve the rates of performance, as the appearance of skin color is a frequent event. The *GetOutCar* class involves a very assorted scenario in terms of color, among other aspects, which makes it more confusing than helpful for the feature representation. However, color information was apparently a decisive factor to improve performance by using the HueSTIP and the Hue-ColorSTIP.

Indoor environments (like offices, bedrooms, living rooms, dining rooms, kitchen) are what prevail in scenes of the *AnswerPhone*, *SitDown* and *Eat* classes, while outdoor

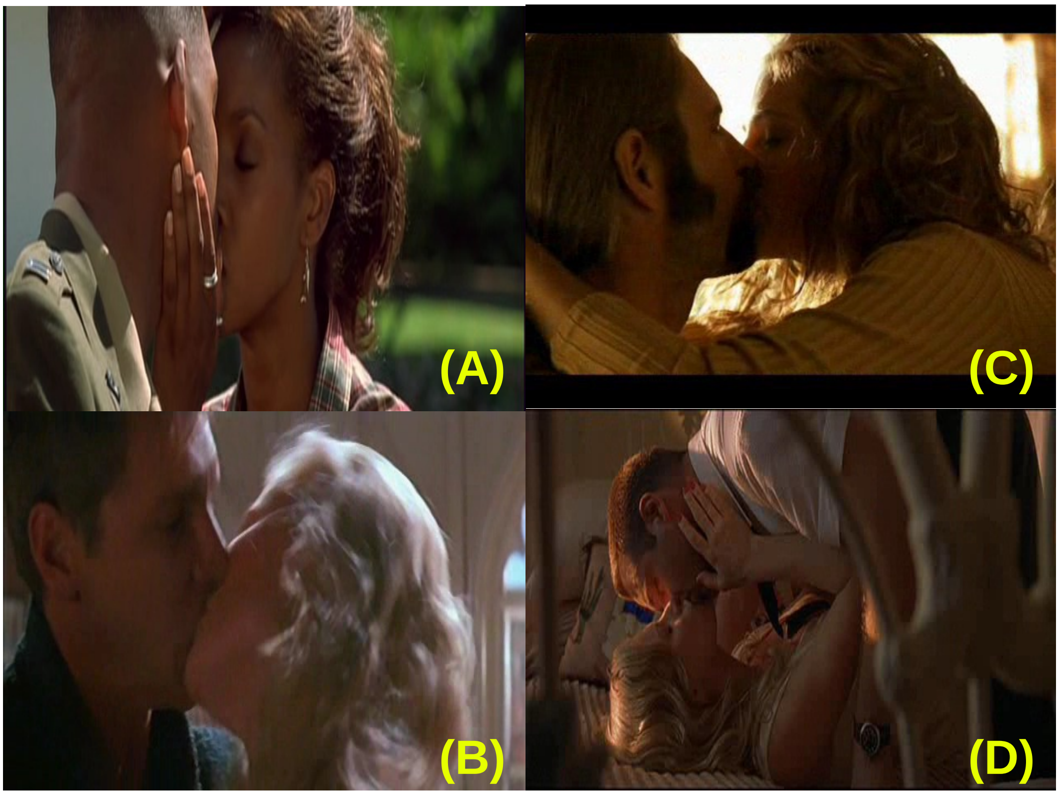

Figure 4.7. Illustration of the set of *Kiss* action videos that were correctly classified by ColorSTIP, but not by STIP

scenes are typical of the *Run* class video clips. Characteristic colors of such scenarios can be helpful to contextualize the detected features defining the actions, which indeed happens to the *AnswerPhone* and *SitDown* classes, but not to the *Eat* and *Run* classes (see Table 4.4). We believe that it was not advantageous to use color-based feature algorithms for recognizing the *Run* class, when grayscale videos are present in the training and test sets. For the *Eat* class, we understood by looking at the confusion matrices (see Tables 4.6 4.7 4.8 4.9) that with the addition of color information at the detection phase the deviations tended to the *Kiss* and *HugPerson*, which can be explained by the similarity in terms of skin color present in the scene.

For the improvements in the *StandUp* class by the color-based versions, we did not identify any evidence supporting the effectiveness of color information. This action and its background context vary a lot, generating noisy features. The same seems to happen for the *SitUp* class, where no improvement was found.

In the results above, it is important to keep in mind the presence of many grayscale video clips, both in the training and the test set, for which color information is simply not available. To evaluate the impact of that missing information on the results, we

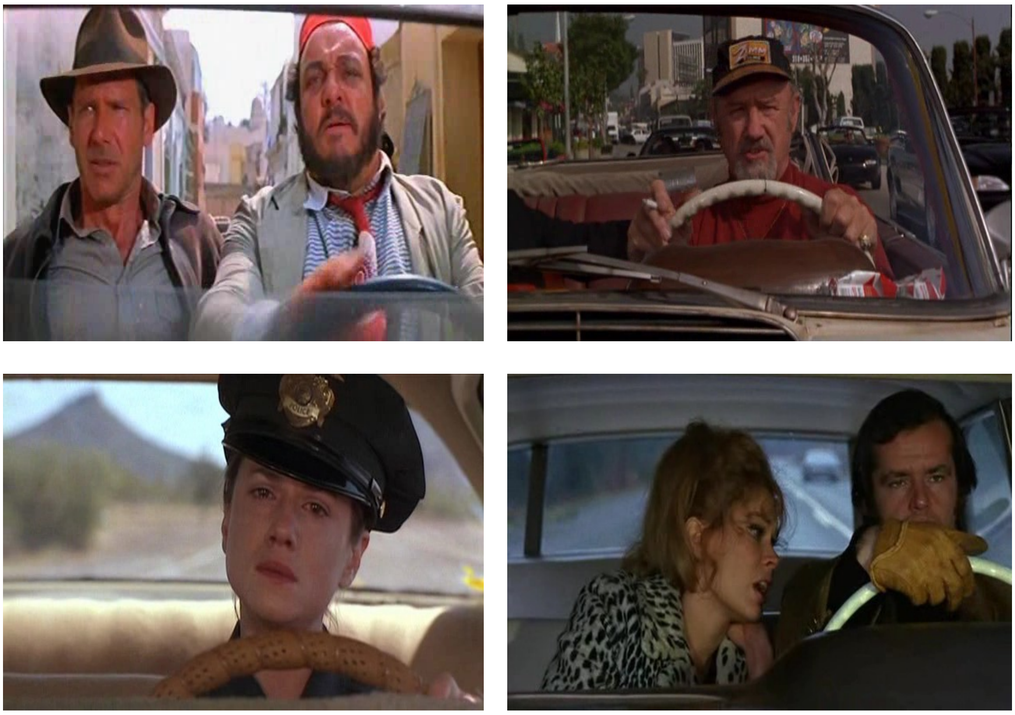

Figure 4.8. Scenes of the *DriveCar* class in which the the camera focuses on the faces of the actors.

repeated the experiment, by removing those videos (Table 4.5).

After we used the color-only dataset to evaluate the family of STIPs, we were able to verify how much consistent the improvements were by comparing with results from the complete dataset. We confirmed that for the classes where color is visually a distinguishable element to define the action (*AnswerPhone*, *DriveCar*, *HandShake*, *HugPerson*, *SitDown*), be in terms of scenario context or of describing the main object parts in performing the action, there were real ameliorations of performance. However, this was not true for the *Kiss* class. We also verified that for the cases where improvement by color information is not intuitive (the *SitUp* and *StandUp* classes), the color-based feature algorithms were worst in performance in comparison with the original STIP. We observed that the color-based algorithms do not worth the computational effort for the *Eat* and *Run* classes, and that for some cases in which the usefulness of color is not obvious the improvements were mantained (the *FightPerson* and *GetOutCar* classes). In this analysis, when color was referred as a useful element for defining the class, many times it was implicit that color would only be useful if accompanied with another type of feature, such as orientation gradients or optic flow.

Table 4.4. This table reports the accuracy rates achieved by each version of the STIP on the Hollywood2 dataset (a multi-class dataset). The best overall performance for each version was chosen to show the results discriminated by classes.

Action	STIP	HueSTIP	ColorSTIP	Hue-ColorSTIP
AnswerPhone	9.4%	7.8%	**12.5%**	10.9%
DriveCar	71.6%	74.5%	**75.5%**	72.6%
Eat	**57.6%**	48.5%	33.3%	39.4%
FightPerson	64.3%	**67.1%**	**67.1%**	62.9%
GetOutCar	15.8%	**21.1%**	15.8%	19.3%
HandShake	6.7%	8.9%	13.3%	**15.6%**
HugPerson	**33.3%**	21.2%	21.2%	19.7%
Kiss	28.2%	37.9%	**50.5%**	45.6%
Run	**58.9%**	56.0%	56.0%	56.7%
SitDown	43.5%	45.4%	**50.9%**	48.2%
SitUp	**2.7%**	0.0%	**2.7%**	0.0%
StandUp	51.4%	61.0%	**63.0%**	58.2%
Average	36.9%	37.4%	**38.5%**	37.4%

4.3.2 Filtering of Unwanted Content: Pornography and Violence

In [Valle *et al.* , 2011], we tested the late fusion combination of features to tackle the problem of recognizing pornographic content in multimedia data. Our proposals were towards application contexts requiring the filtering of unwanted content. We demonstrated that the use of space-time local features (with STIP) are critical to obtain optimal results on this task, both with respect to high true positive rates and lower false positive rates. However, its combination with others (via late fusion) did not significantly improve the number of true positives and slightly increased the false positive rate. We emphasized that STIP was superior for making less mistakes in terms of false negatives. This way, STIP was arguably deemed as the best for the purpose (for more details comparing STIP with other detectors and descriptors in this context, we refer the reader to [Valle *et al.* , 2011]).

In this dissertation, we address the same problem, but by early fusion of features. We embed color information into STIP for detecting and describing color-based space-time interest points, thus creating three extensions of it. We run the experimental protocol described in Section 4.2.4.2 for evaluating and comparing the feature algorithms. In Table 4.13, the results of recall, precision and false positive rate are

Table 4.5. Accuracy rates of the different feature algorithms on the Hollywood2 dataset containing only colorful videos.

Action	STIP	HueSTIP	ColorSTIP	Hue-ColorSTIP
AnswerPhone	6.4%	6.4%	8.5%	**10.9%**
DriveCar	71.1%	74.4%	75.6%	**77.8%**
Eat	**53.9%**	**53.9%**	50.0%	50.0%
FightPerson	63.1%	64.6%	**75.4%**	70.1%
GetOutCar	10.0%	14.0%	**20.0%**	14.0%
HandShake	3.1%	3.1%	**9.4%**	**9.4%**
HugPerson	15.2%	**23.9%**	21.7%	**23.9%**
Kiss	**54.9%**	47.9%	40.9%	39.4%
Run	60.2%	61.2%	**62.2%**	59.2%
SitDown	28.0%	42.7%	**49.3%**	46.7%
SitUp	0.0%	**3.7%**	0.0%	0.0%
StandUp	**65.3%**	52.6%	59.0%	59.0%
Average	35.9%	37.4%	**39.3%**	38.4%

reported, and in Tables 4.11 and 4.12 the confusion matrices are presented. We apply the pairwise t-test to estimate which algorithm is better regarding the true positive rate, false positive rate (FPR) and precision (the results are shown in Table 4.10).

By observing the confusion matrices in Table 4.11, we note that among the color-based algorithms ColorSTIP was the only one to fulfill the requirement of improving the true positive rates, accordingly reducing the rate of false negatives. For applications requiring the removal of abusive content, low rates of false negatives are more valuable. This is either because the interested audience does have strong moral reasons or because the target people are more sensitive. In other contexts, such as the case of forensics systems, false negatives are even more critical. It is desirable to extract the maximum number of suspect data (meaning the necessity of a higher recall), even if false positives are retrieved, provided that the average precision is high. False positives can be tolerated for applications of this nature.In this context, ColorSTIP is definitely the best option.

Note that in Case 1 of ColorSTIP mistakes (see Figure i of the supplemental material [1]), the actors are dressed during the whole film and in Case 2 only the woman's head is displayed in great part of the video. Other common error of ColorSTIP involved back story scenes in a big fraction of the videos. Most of those scenes do not exhibit sexual acts and nudity of the actors (see Figures iv and v). Regarding STIP, a great

[1] As the content of the images may shock some sensibilities, we have made them available as a separate supplement. Available at http://www.npdi.dcc.ufmg.br/colorbasedstips/supplement.pdf

Table 4.6. Matrix confusion evaluating STIP on the Hollywood2 dataset.

Actions	*AnswerPhone*	*DriveCar*	*Eat*	*FightPerson*	*GetOutCar*	*HandShake*	*HugPerson*	*Kiss*	*Run*	*SitDown*	*SitUp*	*StandUp*	***Average***
AnswerPhone	.09	.00	.08	.02	.00	.00	.06	.16	.05	.27	.00	.28	
DriveCar	.00	.72	.04	.09	.02	.00	.04	.02	.05	.00	.00	.03	
Eat	.03	.00	.58	.00	.03	.00	.03	.12	.00	.12	.00	.09	
FightPerson	.01	.03	.00	.64	.01	.00	.03	.03	.20	.01	.00	.03	
GetOutCar	.07	.11	.02	.02	.16	.00	.02	.11	.21	.09	.02	.19	
HandShake	.11	.00	.02	.00	.02	.07	.13	.18	.02	.24	.00	.20	
HugPerson	.03	.00	.03	.12	.02	.06	.33	.17	.05	.09	.02	.09	
Kiss	.09	.02	.06	.05	.04	.02	.20	.28	.03	.11	.01	.10	
Run	.03	.04	.02	.06	.03	.01	.01	.01	.59	.08	.00	.13	
SitDown	.02	.01	.04	.01	.00	.02	.03	.13	.08	.44	.00	.23	
SitUp	.08	.00	.00	.00	.03	.00	.05	.30	.03	.16	.03	.32	
StandUp	.09	.01	.01	.00	.03	.00	.01	.09	.09	.14	.01	.51	
Average													.37

portion of its errors that differed from ColorSTIP's consisted of videos displaying clear scenes of explicit sex. Illustrations from Figures ii and iii of the supplemental material show visible examples of such mistakes. Another interesting case that we observed by analyzing all folds was that, unlike STIP, ColorSTIP hit all cases of pornographic content from cartoons.

We believe that the use of color for describing the motion patterns (HueSTIP) has slightly ameriolated the results for the difficult nonpornographic cases, but the same way color fostered the representation for the challenging cases (see Section 4.1.2 to review about those situations), it weakened (or turned ambiguous) real pornographic videos with similar content to the nonporn hard cases. However, interest points by color-dependent variations in space and time (ColorSTIP) gave a more powerful representation for the pornographic content. This was so expressive that it failed to one case of difficulty (Figure 4.9).

In the context of detecting violent scenes (see Table 4.12), however, STIP excels before its counterpart HueSTIP. Its recall average was superior to HueSTIP's, while the mean value of false positive rates was inferior. This is not so surprising considering that we had already proven in previous work [Valle *et al.* , 2011] that STIP was self-sufficient for distinguishing regular from aggressive events.

To sum up, STIP provided the best characterisitcs for our test on violence de-

Table 4.7. Matrix confusion evaluating HueSTIP on the Hollywood2 dataset.

Actions	*AnswerPhone*	*DriveCar*	*Eat*	*FightPerson*	*GetOutCar*	*HandShake*	*HugPerson*	*Kiss*	*Run*	*SitDown*	*SitUp*	*StandUp*	***Average***
AnswerPhone	.08	.00	.08	.00	.02	.03	.03	.09	.05	.30	.00	.33	
DriveCar	.02	.75	.03	.04	.01	.00	.03	.01	.06	.03	.00	.03	
Eat	.00	.00	.48	.00	.00	.00	.09	.09	.03	.18	.00	.12	
FightPerson	.00	.01	.00	.67	.01	.00	.01	.04	.20	.01	.00	.03	
GetOutCar	.05	.05	.04	.02	.21	.00	.02	.07	.12	.09	.02	.32	
HandShake	.11	.00	.04	.02	.02	.09	.07	.22	.02	.18	.00	.22	
HugPerson	.03	.00	.03	.12	.02	.06	.21	.23	.03	.09	.02	.17	
Kiss	.09	.01	.06	.05	.02	.02	.11	.38	.02	.12	.01	.13	
Run	.02	.03	.01	.06	.02	.01	.00	.03	.56	.07	.00	.19	
SitDown	.02	.00	.00	.01	.00	.01	.03	.11	.05	.45	.00	.32	
SitUp	.14	.00	.00	.00	.03	.00	.03	.24	.05	.11	.00	.41	
StandUp	.05	.00	.01	.00	.03	.01	.01	.11	.07	.10	.00	.61	
Average													.37

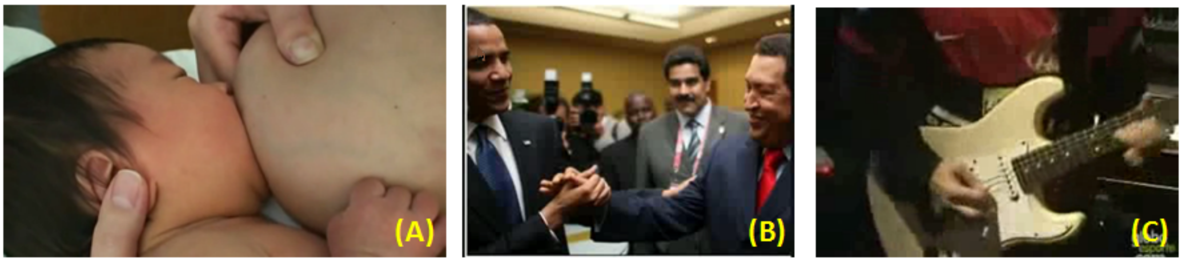

Figure 4.9. Illustrative cases of COLORSTIP false positives, in a round where STIP made no mistakes with respect to the false positive cases.

tection, but it is still disappointing that the early fusion provides no improvements. We believe that in case of violence events, the additional information provided by color is not as clearly associated to specific actions as it seems to happen to pornography. Indeed, color information from videos depicting aggressive acts does not give any differential feature for the action type. The people involved in the scenes are generally dressed, so that not much exposure of skin color is available. Furthermore, the background scenarios vary a lot in terms of color, which visually gives more muddling than deciding clues about the content (see Figure 4.3 for an illustration of those facts).

This means that it is possible to save computational time in terms of image processing, while keeping good rates of hits. Color information was useful, however, in

Table 4.8. Matrix confusion evaluating ColorSTIP on the Hollywood2 dataset.

Actions	*AnswerPhone*	*DriveCar*	*Eat*	*FightPerson*	*GetOutCar*	*HandShake*	*HugPerson*	*Kiss*	*Run*	*SitDown*	*SitUp*	*StandUp*	***Average***
AnswerPhone	.13	.00	.00	.00	.02	.02	.00	.11	.03	.30	.00	.41	
DriveCar	.01	.75	.02	.06	.03	.00	.00	.02	.04	.00	.00	.07	
Eat	.09	.00	.33	.00	.00	.00	.06	.24	.00	.09	.00	.18	
FightPerson	.00	.03	.00	.67	.00	.00	.03	.01	.19	.03	.00	.04	
GetOutCar	.04	.04	.02	.02	.16	.00	.04	.00	.33	.04	.02	.32	
HandShake	.20	.00	.00	.00	.02	.13	.04	.24	.00	.18	.00	.18	
HugPerson	.02	.00	.00	.09	.05	.05	.21	.35	.05	.05	.00	.15	
Kiss	.07	.06	.03	.04	.03	.04	.05	.50	.00	.07	.00	.12	
Run	.03	.03	.02	.09	.05	.01	.01	.01	.56	.05	.00	.14	
SitDown	.07	.02	.00	.02	.01	.04	.01	.07	.06	.51	.00	.19	
SitUp	.16	.00	.00	.00	.03	.03	.05	.11	.05	.11	.03	.43	
StandUp	.07	.01	.00	.01	.02	.01	.01	.07	.06	.10	.01	.63	
Average													.38

our tests to detect pornographic content. Due to the limitations of our scope, we have explored only the contributions of visual information. We understand, however, that systems for those applications might enormously benefit by multimodal schemes, in which visual, auditory, textual and even social information (extracted from the social network data) are synergistically exploited.

Table 4.9. Matrix confusion evaluating Hue-ColorSTIP on the Hollywood2 dataset.

Actions	*AnswerPhone*	*DriveCar*	*Eat*	*FightPerson*	*GetOutCar*	*HandShake*	*HugPerson*	*Kiss*	*Run*	*SitDown*	*SitUp*	*StandUp*	***Average***
AnswerPhone	.11	.00	.00	.00	.02	.02	.03	.17	.03	.28	.00	.34	
DriveCar	.03	.73	.01	.04	.05	.01	.01	.02	.06	.00	.00	.05	
Eat	.09	.00	.39	.00	.00	.00	.15	.12	.00	.09	.00	.15	
FightPerson	.03	.06	.01	.63	.01	.01	.03	.03	.16	.00	.00	.03	
GetOutCar	.04	.05	.04	.00	.19	.00	.05	.00	.32	.05	.00	.26	
HandShake	.24	.00	.00	.00	.00	.16	.00	.20	.00	.18	.00	.22	
HugPerson	.00	.00	.02	.12	.06	.05	.20	.32	.05	.06	.00	.14	
Kiss	.05	.05	.04	.06	.02	.05	.08	.46	.00	.10	.00	.11	
Run	.01	.04	.02	.09	.06	.01	.02	.01	.57	.05	.00	.13	
SitDown	.08	.02	.01	.01	.03	.05	.03	.04	.06	.48	.00	.20	
SitUp	.11	.11	.00	.00	.05	.03	.03	.14	.03	.03	.00	.49	
StandUp	.08	.00	.00	.02	.02	.02	.03	.06	.06	.12	.01	.58	
Average													.37

Table 4.10. Confidence intervals from the pairwise t-test on the pornography and violence datasets, comparing STIP with its color-based extensions.

		Intervals of the Paired Test		
Problem	***Feature***	***Recall***	***Precision***	***FPR***
Porn	STIP X HUESTIP	(-0.005;0.025)	(-0.049;0.023)	(-0.022;0.052)
	STIP X COLORSTIP	**(-0.053;-0.007)**	(-0.042;0.049)	(-0.051;0.041)
Violence	STIP X HUESTIP	(-0.032;0.052)	(-0.001;0.013)	(-0.017;0.007)

Table 4.11. Confusion matrices for the pornography case.

STIP		
Class	*Porn*	*NonPorn*
Porn	**.87**	.13
NonPorn	.075	**.925**

COLORSTIP		
Class	*Porn*	*NonPorn*
Porn	**.90**	.10
NonPorn	.08	**.92**

HUESTIP		
Class	*Porn*	*NonPorn*
Porn	**.86**	.14
NonPorn	.06	**.94**

Table 4.12. Confusion matrices for the violence case

STIP			**HueSTIP**		
Class	*Violence*	*NonViolence*	**Class**	*Violence*	*NonViolence*
Violence	**.91**	.09	*Violence*	**.90**	.10
NonViolence	.15	**.85**	*NonViolence*	.155	**.845**

Table 4.13. Results of the 5-fold cross-validation rounds on the Pornography and Violence dataset.

	Porn - Recall			**Violence - Recall**	
Round	*STIP*	*HueSTIP*	*ColorSTIP*	*STIP*	*HueSTIP*
0	.875	.850	.900	.900	.900
1	.900	.900	.925	.975	.950
2	.850	.850	.850	.850	.900
3	.900	.900	.950	.925	.875
4	.825	.800	.875	.900	.875
Average	.870	.860	0.900	.910	.900

	Porn - Precision			**Violence - Precision**	
Round	*STIP*	*HueSTIP*	*ColorSTIP*	*STIP*	*HueSTIP*
0	.875	.919	.900	.837	.837
1	.947	.947	.949	.830	.826
2	.895	.895	.895	.895	.878
3	1.	.973	.927	.861	.854
4	.892	.942	.921	.878	.875
Average	.922	.935	.918	.860	.854

	Porn - FPR			**Violence -** FPR	
Round	*STIP*	*HueSTIP*	*ColorSTIP*	*STIP*	*HueSTIP*
0	.125	.075	.100	.175	.175
1	.050	.050	.050	.200	.200
2	.100	.100	.100	.100	.125
3	.000	.025	.075	.150	.150
4	.100	.050	.075	.125	.125
Average	.075	.060	0.080	.150	.155

Chapter 5

Conclusion

In this dissertation, we have studied the impact of color-based motion patterns on the classification of actions. Color information was embedded in the STIP algorithm for detection and description of motion patterns. On this basis, we derived three color-based algorithms for extraction of spatiotemporal features (HueSTIP, ColorSTIP and Hue-ColorSTIP). The key idea was to enrich the spatiotemporal detector/descriptor with color information without losing important photometric invariance properties. For this purpose we have employed a well established color invariance model [Gevers *et al.* , 2006]. Two different models of color invariants were used in attempt of enhancing the power of spatiotemporal features, such that they were less affected by the complexity imposed by realistic scenarios and better classification rates could be reached. Color invariants aimed at making our color spatiotemporal features robust to various illumination effects, such as highlights.

All algorithms, including the original STIP, had their performance evaluated under different application contexts, namely content-based video annotation and retrieval, as well as filtering and detection of inappropriate content (pornography and violence). In addition, all experiments were carried out on datasets containing videos downloaded from real sharing social networks, like *YouTube*, and cut off of Hollywood movies.

Experimental results have demonstrated that a color-based STIP can be a better alternative to a few situations, mainly on those cases where color is notably an important aspect to describe the content. As an example, we experienced one important case in which color information did help as much as we expected, the pornography case. At last, after the different experiments on datasets of realistic videos and evaluation on a few in-vogue application contexts, we understand that STIP is still a self-contained tool for solving many real problems at different levels of difficulties.

Future directions of this work lie in

1. Exploring or develop other means of incorporating color information into STIP [Laptev, 2005], or other spatiotemporal feature extractors (e.g., Dollár *et al.* [2005]; Willems *et al.* [2008]);
2. Trying out other color invariant models (which can found in [Gevers *et al.* , 2006]), including the quasi-invariants [van de Weijer *et al.* , 2005];
3. Evaluating the influence of color when using the space-time multichannel approach described in [Laptev *et al.* , 2008b].

As Social Networks evolve, the need to provide tools for semantic classification and retrieval (including to control the proliferation of abusive content) becomes a critical issue. In addition, Digital Forensics has recently arisen as an exciting field of research, which can benefit from our tools to gather evidence from confiscated computers and hard disks, in cases of suspected child pornography, for example. In the latter case experts' time can be dramatically saved by preselecting among hundreds of thousands of documents those which should receive attention. Those and many other applications could take advantage of the proposed contributions.

Publications

As part of this work, we have published the following paper:

[de Souza *et al.* , 2010] de Souza, Fillipe D. M., Chavez, Guillermo C., do Valle Jr., Eduardo A., de A. Araujo, Arnaldo. 2010. *Violence Detection in Video Using Spatio-Temporal Features.* Graphics, Patterns and Images, SIBGRAPI Conference on, 0, 224–230.

We have also produced other works, which are available under the following references:

[Valle *et al.* , 2011] Valle, Eduardo, de Avila, Sandra E. F., da Luz Jr., Antonio, de Souza, Fillipe D. M., de M. Coelho, Marcelo, de Albuquerque Araújo, Arnaldo. 2011. *Content-Based Filtering for Video Sharing Social Networks.* CoRR, abs/1101.2427;

[de Souza *et al.* , 2011] de Souza, Fillipe D. M., Valle, Eduardo, Chávez, Guillermo Cámara, de Albuquerque Araújo, Arnaldo. 2011. *Hue Histograms to Spatiotemporal Local Features for Action Recognition.* CoRR, abs/1104.3742.

Bibliography

Abdelkader, Mohamed F., Roy-Chowdhury, Amit K., Chellappa, Rama, & Akdemir, Umut. 2008. Activity representation using 3D shape models. *J. Image Video Process.*, 2008(January), 5:1–5:16.

Agarwal, S., & Roth, D. 2002. Learning a sparse representation for object detection. *Pages 113–130 of: European Conference on Computer Vision*, vol. 4.

Anandan, P. 1989. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2, 283–310.

Barron, J. L., Fleet, D. J., Beauchemin, S. S., & Burkitt, T. A. 1994. Performance Of Optical Flow Techniques. *International Journal of Computer Vision*, 12(1), 43–77.

Beaudet, P. R. 1978 (Nov.). Rotationally invariant image operators. *Pages 579–583 of: Proceedings of the 4th International Joint Conference on Pattern Recognition.*

Biederman, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol Rev*, 94(2), 115–147.

Black, M. J., Sapiro, G., Marimont, D. H., & Heeger, D. 1998. Robust anisotropic diffusion. *Image Processing, IEEE Transactions on*, 7(3), 421–432.

Bobick, Aaron F., & Davis, James W. 2001. The Recognition of Human Movement Using Temporal Templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(March), 257–267.

Bobick, Aaron F., Intille, Stephen S., Davis, James W., Baird, Freedom, Pinhanez, Claudio S., Campbell, Lee W., Ivanov, Yuri A., Schütte, Arjan, & Wilson, Andrew. 1999. The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. *Presence: Teleoper. Virtual Environ.*, 8(August), 369–393.

Brand, Matthew. 1997. The "Inverse hollywood problem": from video to scripts and storyboards via causal analysis. *Pages 132–137 of: Proceedings of the fourteenth* *national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence.* AAAI'97/IAAI'97. AAAI Press.

Bregler, Christoph. 1997. Learning and Recognizing Human Dynamics in Video Sequences. *Pages 568– of: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97).* CVPR '97. Washington, DC, USA: IEEE Computer Society.

Cedras, Claudette, & Shah, Mubarak. 1995. Motion-Based Recognition: A Survey. *Image and Vision Computing*, 13, 129–155.

Chang, Chih-Chung, & Lin, Chih-Jen. 2001. *LIBSVM: a library for support vector machines.* Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chomat, Olivier, & Crowley, James L. 1999. Probabilistic Recognition of Activity using Local Appearance. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2, 2104.

Cuntoor, Naresh P. 2006. Morse Functions for Activity Classification Using Spatiotemporal Volumes. *Pages 20– of: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop.* CVPRW '06. Washington, DC, USA: IEEE Computer Society.

Cutting, J., Proffitt, Dennis R., & Kozlowski, Lynn T. 1978. A Biomechanical Invariant for Gait Perception. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 357–372.

Dalal, Navneet, & Triggs, Bill. 2005 (June). Histograms of Oriented Gradients for Human Detection. *Pages 886–893 of:* Schmid, Cordelia, Soatto, Stefano, & Tomasi, Carlo (eds), *International Conference on Computer Vision & Pattern Recognition*, vol. 2.

David J., Heeger. 1988. Optical flow using spatiotemporal filters. *International Journal of Computer Vision*, 1, 279–302.

de Souza, Fillipe D. M., Chavez, Guillermo C., do Valle Jr., Eduardo A., & de A. Araujo, Arnaldo. 2010. Violence Detection in Video Using Spatio-Temporal Features. *Graphics, Patterns and Images, SIBGRAPI Conference on*, 0, 224–230.

de Souza, Fillipe Dias Moreira, Valle, Eduardo, Chavez, Guillermo Camara, & de Albuquerque Araújo, Arnaldo. 2011. Hue Histograms to Spatiotemporal Local Features for Action Recognition. *CoRR*, abs/1104.3742.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. 2005 (October). Behavior Recognition via Sparse Spatio-Temporal Features. *In: VS-PETS.*

Doretto, G., Cremers, D., Favaro, P., & Soatto, S. 2003 (October). Dynamic texture segmentation. *Pages 1236–1242 of: Proceedings of the International Conference on Computer Vision*, vol. 2.

Efros, Alexei A., Berg, Alexander C., Mori, Greg, & Malik, Jitendra. 2003. Recognizing Action at a Distance. *In: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision.* Washington, DC, USA: IEEE Computer Society.

Engel, S., & Rubin, J. 1986. Detecting visual motion boundaries. vol. 2.

Fablet, Ronan, Bouthemy, Patrick, & Pérez, Patrick. 2002. Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Transactions on Image Processing*, 11, 393–407.

Fathi, A., & Mori, G. 2008. Action Recognition by Learning Mid-level Motion Features. *In: CVPR.*

Fergus, R., Perona, P., & Zisserman, A. 2003. Object class recognition by unsupervised scale-invariant learning. *Pages 264–271 of: IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2.

Filipovych, R., & Ribeiro, E. 2008. Learning human motion models from unsegmented videos. *Pages 1–7 of: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.*

Fleet, David J., & Jepson, A. D. 1990. Computation of component image velocity from local phase information. *Int. J. Comput. Vision*, 5(September), 77–104.

Forstner, W., & Gulch, E. 1987. A fast operator for detection and precise location of distinct points, corners and centres of circular features. 281–305.

Förstner, Wolfgang. 1994. A framework for low level feature extraction. *Pages 383–394 of: Proceedings of the third European conference on Computer Vision (Vol. II).* Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Gaitanis, Kosta, Correa, Pedro, & Macq, Benoit M. 2006. Modelization of Limb Coordination for Human Action Analysis. *Pages 1765–1768 of: ICIP.*

Gavrila, D. M., & Davis, L. S. 1995. Towards 3-D model-based tracking and recognition of human movement: a multi-view approach. *Pages 272–277 of: In International Workshop on Automatic Face- and Gesture-Recognition. IEEE Computer Society.*

Geusebroek, Jan-Mark, van den Boomgaard, Rein, Smeulders, Arnold W.M., & Geerts, Hugo. 2001. Color Invariance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(December), 1338–1350.

Gevers, T., & Stokman, H. 2004. Robust Histogram Construction from Color Invariants for Object Recognition. *PAMI*, 26(January), 113–117.

Gevers, Th., van de Weijer, J., & Stokman, H. 2006. *Color Feature Detection: An Overview.* R. lukac and k.n. plataniotis (eds.): Color image processing: Methods and applications edn. CRC Press.

Giraudon, G., & Deriche, R. 1991. Accurate Corner Detection: An Analytical Study.

Gorelick, Lena, Blank, Moshe, Shechtman, Eli, Irani, Michal, & Basri, Ronen. 2007. Actions as Space-Time Shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12), 2247–2253.

Gould, K., & Shah, M. 1989. The Trajectory Primal Sketch: A Multi-Scale Scheme For Representing Motion Characteristics. *CVPR*, 89, 79–85.

Guiducci, A. 1988. Corner characterization by differential geometry techniques. *Pattern Recogn. Lett.*, 8(December), 311–318.

Harris, C., & Stephens, M.J. 1988. A Combined Corner and Edge Detector. *Pages 147–152 of: Alvey Vision Conference.*

Heeger, David J. 1987. Model for the extraction of image flow. *J. Opt. Soc. Am. A*, 4(8), 1455–1471.

Hoey, Jesse, & Little, James J. 2003. Bayesian Clustering of Optical Flow Fields. *Pages 1086– of: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2.* ICCV '03. Washington, DC, USA: IEEE Computer Society.

Horn, Berthold K.P., & Schunck, Brian G. 1980. *Determining Optical Flow.* Tech. rept. Cambridge, MA, USA.

Hsu, Chih-Wei, & Lin, Chih-Jen. 2002. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415–425.

Hu, M., Ali, S., & Shah, M. 2008. Learning Motion Patterns in Crowded Scenes Using Motion Flow Field. *In: Proc. International Conference on Pattern Recognition (ICPR).*

Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., & Huang, T.S. 2009. Action Detection in Complex Scenes with Spatial and Temporal Ambiguities. *In: IEEE International Conference on Computer Vision.*

Isard, M., & Blake, A. 1998. Condensation – conditional density propagation for visual tracking.

Jiang, Hao, Drew, Mark S., & Li, Ze-Nian. 2006. Successive Convex Matching for Action Detection. *Pages 1646–1653 of: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2.* CVPR '06. Washington, DC, USA: IEEE Computer Society.

Kadir, T., & Brady, M. 2001. Saliency, scale and image description. *IJCV*, 45(2), 83–105.

Ke, Yan, & Sukthankar, Rahul. 2004. PCA-SIFT: A more distinctive representation for local image descriptors.

Ke, Yan, Sukthankar, Rahul, & Hebert, Martial. 2007. Spatio-temporal Shape and Flow Correlation for Action Recognition. *In: CVPR.*

Kitchen, Les, & Rosenfeld, Azriel. 1982. Gray-level corner detectionâ. *Pattern Recognition Letters*, 1, 95–102.

Kläser, Alexander, Marszałek, Marcin, & Schmid, Cordelia. 2008 (sep). A Spatio-Temporal Descriptor Based on 3D-Gradients. *Pages 995–1004 of: British Machine Vision Conference.*

Klinker, G., Shafer, Steven, & Kanade, Takeo. 1991. Using a Color Reflection Model to Separate Highlights from Object Color. *Pages 145–150 of: Proc. First International Conference on Computer Vision.*

Klinker, Gudrun J., Shafer, Steven A., & Kanade, Takeo. 1989. A physical approach to color image understanding. *Int. J. Comput. Vision*, 4(December), 7–38.

Laptev, I. 2005. On Space-Time Interest Points. *IJCV*, 64(2-3), 107–123.

Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B. 2008a (jun). Learning Realistic Human Actions from Movies. *In: CVPR.*

Laptev, Ivan. 2004. *Local Spatio-Temporal Image Features for Motion Interpretation.* PhD in Computer Science, KTH Numerisk analys och datalogi.

Laptev, Ivan. 2006. Improvements of object detection using boosted histograms. *Pages 949–958 of: In Proc. BMVC.* BMVC.

Laptev, Ivan, & Lindeberg, Tony. 2003. Space-time Interest Points. *Pages 432–439 of: ICCV.*

Laptev, Ivan, & Lindeberg, Tony. 2004. Local descriptors for spatio-temporal recognition. *In: In First International Workshop on Spatial Coherence for Visual Motion Analysis.*

Laptev, Ivan, & Pérez, Patrick. 2007. Retrieving actions in movies. *IEEE International Conference on Computer Vision.*

Laptev, Ivan, Marszałek, Marcin, Schmid, Cordelia, & Rozenfeld, Benjamin. 2008b. Learning Realistic Human Actions from Movies. *In: IEEE Conference on Computer Vision & Pattern Recognition.*

Lazebnik, S., Schmid, C., & Ponce, J. 2004. Semi-local affine parts for object recognition. *Pages 959–968 of: British Machine Vision Conference*, vol. 2.

Lopes, Ana Paula Brandão, do Valle Jr., Eduardo Alves, de Almeida, Jussara Marques, & de Albuquerque Araújo, Arnaldo. 2010. Action Recognition in Videos: from Motion Capture Labs to the Web. *CoRR*, abs/1006.3506.

Lopes, Ana Paula BrandÃ£o, Oliveira, Rodrigo Silva, de Almeida, Jussara Marques, & de Albuquerque AraÃºjo, Arnaldo. 2009. Spatio-Temporal Frames in a Bag-of-Visual-Features Approach for Human Actions Recognition. *Pages 315–321 of: SIBGRAPI.* IEEE Computer Society.

Lowe, David G. 1999. Object Recognition from Local Scale-Invariant Features. *Pages 1150– of: Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2.* ICCV '99. Washington, DC, USA: IEEE Computer Society.

Lowe, David G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2), 91–110.

Lucas, Bruce D., & Kanade, Takeo. 1981. An iterative image registration technique with an application to stereo vision. *Pages 674–679 of: Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Marszałek, Marcin, Laptev, Ivan, & Schmid, Cordelia. 2009. Actions in Context. *In: IEEE Conference on Computer Vision & Pattern Recognition.*

Mikolajczyk, Krystian, & Schmid, Cordelia. 2002. An affine invariant interest point detector. *Pages 128–142 of: Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark.* Springer. Copenhagen.

Mikolajczyk, Krystian, & Schmid, Cordelia. 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10), 1615–1630.

Mokhber, Arash, Achard, Catherine, & Milgram, Maurice. 2008. Recognition of human behavior by space-time silhouette characterization. *Pattern Recogn. Lett.*, 29(January), 81–89.

Mokhtarian, Farzin, & Suomela, Riku. 1998. Robust Image Corner Detection Through Curvature Scale Space. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(December), 1376–1381.

Montesinos, P., Gouet, V., & Deriche, R. 1998. Differential invariants for color images. *Pages 838–840 of: Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol. 1.

Moravec, Hans P. 1981. Rover visual obstacle avoidance. *Pages 785–790 of: Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Nagel, H. 1982. Displacement vectors derived from second order intensity variations in image sequences. *Graphical Models /graphical Models and Image Processing /computer Vision, Graphics, and Image Processing*, 20, 296–296.

Niebles, Juan Carlos, Wang, Hongcheng, & Fei-Fei, Li. 2008. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *International Journal of Computer Vision*, 79(3), 299–318.

Oikonomopoulos, A., Patras, I., & Pantic, M. 2006. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man and Cybernetics - Part B*, 36(3), 710–719.

Polana, R., & Nelson, R. 1997. Temporal texture and activity recognition. *Pages 87–124 of:* Shah, M., & Jain, R. (eds), *Computational Imaging and Vision series*, vol. 9. Kluwer Academic Publishers.

Rapantzikos, K., Avrithis, Y., & Kollias, S. 2009. Dense saliency-based spatiotemporal feature points for action recognition. CVPR.

Rubin, J., & Richards, W. 1985. *Boundaries of visual motion.*

Rui, Yong, & Anandan, P. 2000. Segmenting Visual Actions Based on Spatio-Temporal Motion Patterns. *Pages 1111–1118 of: Computer Vision and Pattern Recognition.*

Sarkar, Sudeep, Phillips, P. Jonathon, Liu, Zongyi, Vega, Isidro Robledo, Grother, Patrick, & Bowyer, Kevin W. 2005. The HumanID Gait Challenge Problem: Data Sets, Performance, and Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2), 162–177.

Schmid, Cordelia, Mohr, Roger, & Bauckhage, Christian. 1998. Comparing and Evaluating Interest Points. *In: Proceedings of the 6th International Conference on Computer Vision, Bombay, India.* IEEE Computer Society Press.

Schmid, Cordelia, Mohr, Roger, & Bauckhage, Christian. 2000. Evaluation of Interest Point Detectors. *International Journal of Computer Vision*, 37(2), 151–172.

Schuldt, Christian, Laptev, Ivan, & Caputo, Barbara. 2004. Recognizing Human Actions: A Local SVM Approach. *Pages 32–36 of: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03.* ICPR '04. Washington, DC, USA: IEEE Computer Society.

Scovanner, Paul, Ali, Saad, & Shah, Mubarak. 2007. A 3-dimensional sift descriptor and its application to action recognition. *Pages 357–360 of: Proceedings of the 15th international conference on Multimedia.* MULTIMEDIA '07. New York, NY, USA: ACM.

Shafer, Steven A. 1985. Using color to separate reflection components. *COLOR research and application*, 10, 210–218.

Sidenbladh, Hedvig, & Black, Michael J. 2001. Learning Image Statistics for Bayesian Tracking. *Pages 709–716 of: In IEEE International Conference on Computer Vision.*

Siebel, Nils T, & Maybank, Stephen J. 2004 (May). The ADVISOR visual surveillance system. *Pages 103–111 of:* Clabian, Markus, Smutny, Vladimir, & Stanke, Gerd (eds), *Proceedings of the ECCV 2004 workshop "Applications of Computer Vision" (ACV'04), Prague, Czech Republic.*

Singh, Ajit, & Allen, Peter. 1992. Image-flow computation: an estimation-theoretic framework and a unified perspective. *CVGIP: Image Underst.*, 56(September), 152–177.

Sminchisescu, Cristian, & Triggs, Bill. 2003. Kinematic jump processes for monocular 3D human tracking. *Pages 69–76 of: Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition.* CVPR'03. Washington, DC, USA: IEEE Computer Society.

Snoek, Cees G. M., Worring, Marcel, & Smeulders, Arnold W. M. 2005. Early versus late fusion in semantic video analysis. *Pages 399–402 of: Proceedings of the 13th annual ACM international conference on Multimedia.* MULTIMEDIA '05. New York, NY, USA: ACM.

Song, Yang, Goncalves, Luis, & Perona, Pietro. 2003. Unsupervised Learning of Human Motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(July), 814–827.

Tuytelaars, Tinne, & Mikolajczyk, Krystian. 2008. *Local Invariant Feature Detectors: A Survey.* Hanover, MA, USA: Now Publishers Inc.

Ullman, Shimon, & Sha'ashua, Amnon. 1988. *Structural Saliency: The Detection of Globally Salient Structures Using a Locally Connected Network.* Tech. rept. AIM-1061.

Uras, Sergio, Girosi, Federico, Verri, Alessandro, & Torre, Vincent. 1988. A computational approach to motion perception. *Biological Cybernetics*, 60, 79–97.

Valle, Eduardo, de Avila, Sandra Eliza Fontes, da Luz Jr., Antonio, de Souza, Fillipe Dias Moreira, de M. Coelho, Marcelo, & de Albuquerque Araújo, Arnaldo. 2011. Content-Based Filtering for Video Sharing Social Networks. *CoRR*, abs/1101.2427.

van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. 2010. Evaluating Color Descriptors for Object and Scene Recognition. *PAMI*, 32(9), 1582–1596.

van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. 2010. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 1582–1596.

van de Weijer, J., Gevers, T., & Geusebroek, J. M. 2005. Edge and Corner Detection by Photometric Quasi-Invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4), 625–630.

van de Weijer, Joost, & Schmid, Cordelia. 2006. Coloring local feature extraction. *Pages 334–348 of: ECCV*, vol. Part II. Springer.

van de Weijer, Joost, Gevers, Theo, & Bagdanov, Andrew D. 2006. Boosting Color Saliency in Image Feature Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(January), 150–156.

van Gemert, J. C., Burghouts, G. J., Seinstra, F. J., & Geusebroek, J. M. 2006. Color Invariant Object Recognition Using Entropic Graphs. *International Journal of Imaging Systems and Technology*, 16(5), 146–153.

Viitaniemi, Ville, & Laaksonen, Jorma. 2008. Experiments on Selection of Codebooks for Local Image Feature Histograms. *Pages 126–137 of: Proceedings of the 10th international conference on Visual Information Systems: Web-Based Visual Information Search and Management.* VISUAL '08. Berlin, Heidelberg: Springer-Verlag.

Viola, Paul A., & Jones, Michael J. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. *Pages 511–518 of: CVPR (1).*

Wallach, H., & O'Connell, D. N. 1953. The kinetic depth effect. *Journal of Experimental Psychology*, 45, 205–217.

Wang, Han, & Brady, Michael. 1995. Real-time corner detection algorithm for motion estimation. *Image Vision Comput.*, 13(9), 695–703.

Wang, Heng, Ullah, Muhammad Muneeb, Kläser, Alexander, Laptev, Ivan, & Schmid, Cordelia. 2009 (sep). Evaluation of local spatio-temporal features for action recognition. *Page 127 of: British Machine Vision Conference.*

Waxmann, A., Wu, J., & Bergholm, F. 1998. Convected activation profiles and receptive fields for real time measurement of short range visual motion. *Pages 717–723 of: Proceedings of IEEE CVPR.*

Weber, M., Welling, M., & Perona, P. 2000. Towards automatic discovery of object categories. *Pages 2101–2109 of: IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2.

Willems, Geert, Tuytelaars, Tinne, & Gool, Luc. 2008. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. *Pages 650–663 of: ECCV*. Berlin, Heidelberg: Springer-Verlag.

Wong, Shu-Fai, & Cipolla, Roberto. 2007. Extracting Spatiotemporal Interest Points using Global Information. *Pages 1–8 of: ICCV.*

Würtz, R.P., & Lourens, T. 1997. *Corner detection in color images by multiscale combination of end-stopped cortical cells.*

Yilmaz, Alper, & Shah, Mubarak. 2005. Actions Sketch: A Novel Action Representation. *Pages 984–989 of: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01.* CVPR '05. Washington, DC, USA: IEEE Computer Society.

YouTube. 2010. *Great Scott! Over 35 Hours of Video Uploaded Every Minute to YouTube.* `http://youtube-global.blogspot.com/2010/11/great-scott-over-35-hours-of-video.html`.

Zelnik-Manor, L., & Irani, M. 2001. Event-based analysis of video. vol. 2.