UNIVERSITÀ
DEGLI STUDI
FIRENZE

Università degli Studi di Firenze
Dipartimento di Ingegneria dell'Informazione (DINFO)
Corso di Dottorato in Ingegneria dell'Informazione
Curriculum: Xxxxxx Xxxxxx

Title of PhD Thesis

Candidate
Name Surname

Supervisors
Prof. XXXXX YYYYY

Dr. XXXXX YYYYYY

PhD Coordinator
Prof. XXXXX YYYYYY

ciclo xxxxx, 20xx-20xx

Università degli Studi di Firenze, Dipartimento di Ingegneria
dell'Informazione (DINFO).

A XXXXX

Acknowledgments

I would like to acknowledge the efforts and input of my supervisor, Prof. XXX YYY, and all my colleagues of the XXXX Lab (XLAB) who were of great help during my research. In particular my thanks go to XXX YYYYY, AAAAAA BBBBB and LLLLL DDDDDDD who collaborated on the main parts of my research work. I would like to thank also Dr. XXXXX XXXXXX for his kind support and hospitality during my stay at XXXXXXX.

Contents

Chapter 1

Introduction

1.1 The objective

1.2 Contributions

Chapter 2

Literature review

This chapter gives a brief survey of related work on object and event recognition using local visual features. The first part of the chapter roughly introduces the problem of object recognition in image archives, while the second part deals with the problem of semantic video annotation. Finally, multimedia ontologies have been presented as a formal tool to enrich the semantic image/video annotation or to derive new knowledge. [1]

[1] The part of this chapter related to semantic video annotation has been published as "Event detection and recognition for semantic annotation of video" in *Multimedia Tools and Applications (Special Issue: Survey Papers in Multimedia by World Experts), vol. 51, iss. 1, pp. 279-302, 2011* [**?**].

2.1 Recognition of object instances

2.1.1 Local visual features

2.2 Recognition of object categories

2.2.1 Codebooks

2.3 Semantic video annotation

2.3.1 Actions and events

2.3.2 Spatio-temporal features

2.3.3 Classification of composite events

2.4 Ontologies

Chapter 3

Trademark retrieval in sports video archives

In this chapter we describe a system for detection and retrieval of trademarks appearing in sports videos. We propose a compact representation of trademarks and video frame content based on SIFT feature points. This representation can be used to robustly detect, localize, and retrieve trademarks as they appear in a variety of different sports video types. Classification of trademarks is performed by matching a set of SIFT feature descriptors for each trademark instance against the set of SIFT features detected in each frame of the video. Localization is performed through robust clustering of matched feature points in the video frame. Experimental results are provided, along with an analysis of the precision and recall. Results show that the our proposed technique is efficient and effectively detects and classifies trademarks. [1] [2]

[1] This chapter has been published as "Trademark Matching and Retrieval in Sports Video Databases" in *Proc. of ACM Multimedia Information Retrieval (MIR), 2007* [**?**].

[2] *Acknowledgments:* this work was partially supported by Sport System Europe srl, Bologna, Italy.

Chapter 4

Context-dependent trademark matching and retrieval

In this chapter we introduce a novel multiple-logo matching and detection algorithm based on a new class of similarity functions referred to as context dependent. Our approach is based on designing a similarity measure, involving interest points, which takes into account not only their intrinsic visual features but also their context and spatial configuration. The main contribution of this work includes (i) a variational framework which makes it possible to design our similarity as the fixed point of an energy function mixing a visual "data term", a "context criterion" and a "regularization term" and, (ii) a theoretical study of the consistency of logo matching/detection and its invariance to different transformations including similarity and occlusion. Finally, we will show the validity of the method through extensive experiments on challenging logo images. [1]

[1]Part of this work was conducted while the author was a visiting Ph.D. student at XXXXX XXXXX, City (Nation), from Month to Month YEAR (working with. Dr. Xxxx Yyyy). This chapter previously appeared as research report n. 2010D009, Xxxxx XXX [**?**].

Chapter 5

A SIFT-based forensic method for copy-move detection

One of the principal problem image forensics has to deal with is determining if a particular image is authentic or not. This task is very important in all those fields where is crucial to use such digital content as evidence like, for instance, in a court of law. To carry out such forensic analysis various technological instruments have been developed in literature. Many of them try to reveal if some modifications have been performed thus assessing that something of suspect could have been made, other ones search for comprehending what has happened and possibly which relations there are with other linked photos. In this chapter the problem of detecting if a feigned image has been created is investigated; in particular, attention has been paid to the case in which an area of an image is copied and then pasted onto another zone to make a duplication or to cancel something that was awkward. To detect such modifications we propose a new methodology based on SIFT features. Our method allows both to understand if a copy-move attack has occurred and which are the image points involved, and, furthermore, to recover which has been the geometric transformation happened to perform cloning. [1]

[1] A preliminary version of the work presented in this chapter has been published as "Geometric tampering estimation by means of a SIFT-based forensic analysis" in *Proc. of IEEE Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010*, [**?**]

5.1 Introduction

5.2 SIFT Features for Image Forensics

5.2.1 Our contribution

5.3 The proposed method

5.3.1 SIFT features extraction and multiple keypoint matching

5.3.2 Clustering and forgeries detection

5.3.3 Geometric transformation estimation

5.4 Experimental results

5.4.1 Settings for forgery detection

5.4.2 Test on multiple copied regions

5.4.3 Test on a large dataset

5.4.4 Image splicing

5.5 Conclusion

and submitted to IEEE Trans. on Information Forensics and Security (TIFS).

Chapter 6

Video event classification using string kernels

Event recognition is a crucial task to provide high-level semantic description of the video content. The bag-of-words (BoW) approach has proven to be successful for the categorization of objects and scenes in images, but it is unable to model temporal information between consecutive frames. In this chapter we present a method to introduce temporal information for video event recognition within the BoW approach. Events are modeled as a sequence composed of histograms of visual features, computed from each frame using the traditional BoW. The sequences are treated as strings (phrases) where each histogram is considered as a character. Event classification of these sequences of variable length, depending on the duration of the video clips, are performed using SVM classifiers with a string kernel that uses the Needlemann-Wunsch edit distance. Experimental results, performed on two domains, soccer videos and a subset of TRECVID 2005 news videos, demonstrate the validity of the proposed approach. [1]

[1] This chapter has been published as "Video Event Classification using String Kernels" in *Multimedia Tools and Applications, vol. 48, iss. 1, pp. 69-87, 2010* [?].

Chapter 7

Effective codebooks for human action categorization

In this chapter we propose a new method for human action categorization, combining novel gradient and optic flow descriptors, and creating an effective bag-of-words model. Recent approaches have represented videos using bag of spatio-temporal visual words, following the successful results achieved in object and scene classification. In such cases codebooks are usually obtained by k-means clustering and hard assignment of visual features to the more representative codewords. Our main contribution is two-fold. First, we define a novel 3D gradient descriptor that combined with optic flow outperforms the state-of-the-art, without requiring fine parameter tuning. Second, we show that for spatio-temporal features the popular k-means algorithm is insufficient, because cluster centers are attracted by the denser regions of the sample distribution, providing a non-uniform description of the feature space and thus failing to code other informative regions. We obtain a more effective codebook by applying a radius-based clustering method and a soft assignment that considers the information of two or more relevant codeword candidates. [1]

[1] A preliminary version of the work presented in this chapter has been published as "Effective Codebooks for Human Action Categorization" in *Proc. of ICCV International Workshop on Video-oriented Object and Event Classification (VOEC), 2009* [**?**].

Chapter 8

Video annotation using ontologies and rule learning

In this chapter we present an approach for automatic annotation and retrieval of video content, based on ontologies and semantic concept classifiers. A novel rule-based method is used to describe and recognize composite concepts and events. Our algorithm learns automatically rules expressed in Semantic Web Rules Language (SWRL), exploiting the knowledge embedded into the ontology. The relationship between concepts, their co-occurrence and the temporal consistency of video data are used to improve the performance of individual concept detectors. Finally, we present a web video search engine, based on ontologies, that permits queries using a composition of boolean and temporal relations between concepts. [1]

[1] This chapter has been published as "Video Annotation and Retrieval Using Ontologies and Rule Learning" in *IEEE MultiMedia, vol. 17, iss. 4, pp. 80-88, 2010* [**?**].

Chapter 9

Conclusion

This chapter summarizes the contribution of the thesis and discusses avenues for future research.

9.1 Summary of contribution

9.2 Directions for future work

Appendix A

Appendix

This appendix is related to XXXXXX, previously presented in Chapter **??**. Here we proof XXXXXX (see Section **??**).

A.1 Proof of proposition 2 in Section ??

Appendix B

Publications

This research activity has led to several publications in international journals and conferences. These are summarized below.[1]

International Journals

1. **Thesis Author**, X. XXXXXX, X. XXXX, X. XXXXXXXX. "Title of journal paper", *Name of Journal*, vol. in press, 201x. (Special Issue: SXXXX) [DOI:10.1007/s11042-010-0643-7]
2. **Thesis Author**, X. XXXXXX, X. XXXX, X. XXXXXXXX. "Title of other journal paper", *IEEE Name of Journal*, vol. XX, iss. X, pp. xx-xx, 201x. [DOI: 10.1109/MMUL.2010.4]

Submitted

1. **Thesis author**, X. XXXXXX, X. XXXX, X. XXXXXXXX. "Hopefully the paper with this title will be accepted", *EName of Journal*, 201x. (Submitted after major revision)

International Conferences and Workshops

1. **Thesis Author**, X. XXXXXX, X. XXXX, X. XXXXXXXX. "Title of conference paper", in *Proc. of ACM International Workshop or Conference on XXXXXX (ACRNYM)*, City (Nation), 201x. **(Best paper award)**

[1] The author's bibliometric indices are the following: H-index = X, total number of citations = XX (source: Google Scholar on Month XX, 201x).

2. **Thesis Author**, X. XXXXXX, X. XXXX, X. XXXXXXXX. "Title of conference paper", in *Proc. of IEEE International Workshop or Conference on XXXXXX (ACRNYM)*, City (Nation), 201x.

National Conferences

1. **Thesis Author**, X. XXXXXX, X. XXXX, X. XXXXXXXX. "TItle of minor paper", in *Proc. of XXXX National Conference*, Location (Province), Italy, 201x.

Technical Reports

1. **Thesis Author**, X. XXXXXX, X. XXXX, X. XXXXXXXX. "Title of Tech report", TELECOM ParisTech, Technical Report, TechReportCode, 201x.