16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

ACTOR RETRIEVAL SYSTEM BASED ON KERNELS ON BAGS OF BAGS*

Shuji Zhao[1], Frédéric Precioso[1], Matthieu Cord[2], Sylvie Philipp-Foliguet[1]

[1]ETIS, CNRS, ENSEA, Univ Cergy-Pontoise
6, avenue du Ponceau, F-95000 Cergy-Pontoise, France
phone: +33 130736288
email: {zhao, precioso, philipp}@ensea.fr

[2]LIP6, UPMC, CNRS
104, avenue du Président Kennedy, F-75016 Paris, France
phone: +33 144277139
email: matthieu.cord@lip6.fr

ABSTRACT

In the domain of multimedia, rapid DVD browsing or multimedia oriented web search require an efficient content-based image and video retrieval system. In this paper, we present our retrieval system of actors in films combining powerful machine learning techniques with "kernels on bags of bags" design. From a film segmented into shots, we extract video-tubes of actor faces and represent these video objects with sets of temporally coherent features. These visual features are then input in the kernel-based SVM retrieval system. Our approach has been tested on retrieving actors in a real film and proved its efficiency.

1. INTRODUCTION

Face recognition is a very active topic in image processing research, even more in the last ten years [15]. Comparing the little amount of data and the few contexts in video databases with the wide range of contexts and data for still image databases [14], regarding all publications in the field, one understands that face recognition methods are mainly image-based and mostly target security and biometrics applications.

However, recent achievements in face detection techniques [7, 10, 11, 12] allowed new aims for face recognition task. Arandjelovic et al. [1] and Sivic et al. [2] consider face recognition as a process to retrieve actors in films sequences. This is also the context of our work.

Arandjelovic et al. [1] propose to retrieve actors in shot sequences using a frame-based description. They extract a "face signature" from each frame of the shot and use the set of signatures in the retrieval process. They first defined a distance between two face signatures, a query and a reference, then extended it to a distance between a query and a reference set of face signatures.

In our approach, we want to take into account the temporal coherency. We propose to consider the spatiotemporal tube, containing several instances (one instance per frame of the shot) of the segmented object of interest (here the face of an actor), identified and tracked, as one video object. In our context, classification and learning methods will be applied on video tubes (spatiotemporal tubes) of detected persons. Sivic et al. in [2] first proposed this concept of video tubes. In order to analyze the visual content, Sivic et al. represent the distribution of features of each tube, over a predefined visual word dictionary, as a single vector. The efficiency of this approach is directly dependant on the relevance of the dictionary.

In the work of Zhang et al [20], the authors describe images with sets of features extracted from sparse keypoint locations. In order to represent the distribution of these sets over the training and test images, they propose two approaches: either a clustering in the feature space to reduce the size of the sets or a projection onto a learnt visual vocabulary.

In this paper, we propose to describe a video tube by a set of sets of local features -- a bag of bags of vectors. Each video tube is represented by a set of temporally consistent chains of local descriptors SIFT (Scale Invariant Feature Transform [9]).

The main contribution of this paper is to introduce a kernel design on such complex objects. This framework will allow us to use powerful statistical learning strategies, which have proved their efficiency on image and video retrieval [4, 6, 8].

A query consists in providing our retrieval system with one or several tubes of the requested person. The system returns the tubes concerning the same person ranked by relevance.

We will describe, in chapter 2, visual feature extraction and tracking. In chapter 3, we will present the machine learning techniques and the design of a dedicated kernel. In chapter 4, we show some retrieval results on an extract of a real film. We then present some perspectives and future work with the conclusion.

2. VIDEO OBJECT EXTRACTION AND DESCRIPTION

2.1 Video Tubes

The existing methods of face detection are fairly mature and have been clearly described in [7, 10, 11, 12]. In our work we use the popular algorithm AdaBoost of [7] extended by Lienhart et al [13] implemented in the OpenCV library to detect frontal faces and profile faces in feature films. In our work we use ellipses (instead of the usual rectangles) to approximate the contour of the face (Fig. 1).

We have used, in this paper, our temporal segmentation algorithm presented in [16]. We extract from each shot containing one face (respectively several faces) the video-tube (respectively the video-tubes) made of the face region in the successive frames (Fig. 3(a)).

Even though the face contour is sometimes coarsely approximated by an ellipse shape, the video-tubes define the basic representation of video objects from now on.

*This work is funded by *K-VideoScan* Digiteo Project N°2007-34HD and *iTOWNS* ANR MDCO 2007 Project.

Figure 1 - Face detection by AdaBoost.
Ellipse in red: frontal face; ellipse in blue: right profile face; ellipse in green: left profile face.

2.2 Visual Feature Extraction

2.2.1 Feature Extraction

Visual feature extraction is based on SIFT points of interest extractor [9]. The SIFT descriptor provides then, for each SIFT point, a local description of the image. The SIFT descriptor has been proved to be robust in face recognition [1].
The SIFT points are automatically detected, for each frame of the tube, and a 128-dimensional SIFT descriptor, representing the 16 8-bin histograms of image gradient orientations inside a 4 x4 spatial grid over the image, is extracted.
Our method is based on a relevant description of face videotubes without introducing any a priori knowledge; feature extraction is not related to any learning process, any face model or generic mesh.
One video object is described by the full set of SIFT vectors extracted in the tube.

2.2.2 Temporal Coherency Filtering

Since we do not pre-process the video tubes, the relevant SIFT points, which are quite stable, present in several successive frames of the tube (for instance SIFT on the nose, on the eyes, on the mouth…), are mixed up with other SIFT points that come from visual artefacts like video compression effects, etc.
In order to keep the most significant points, we filter non-persistent points with a robust tracking strategy based on the expected temporal coherency of visual features in the video tube. Furthermore, a SIFT point is considered as being tracked, as long as its position, its scale and its 128 values remain temporally coherent along the tube. Thus, a tracked point can be seen like defining a chain of points. These chains represent the relevant information support of a face in the tube. Chains of less than, or equal to, 2 points are considered as noise and are removed (Fig. 2).
Our temporal coherency is used to the tracking of SIFT points, instead of the tracking of face region as in [2].
An example of tracking and filtering on five consecutive frames is showed in Fig. 3.
One video object is thus represented by a set of consistent SIFT chains. The number of SIFT points extracted per frame vary; hence a tube is represented by several chains, of points of interest, of varying length.
Next section focuses on the design of a similarity measure between such two actor face tubes.

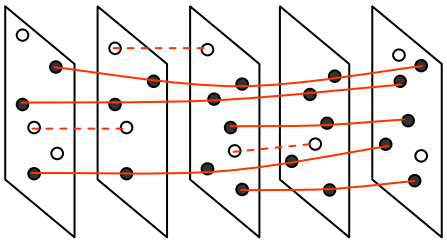

Figure 2 - Chains of tracked points in a tube
(solid lines: persistent points,
dash lines: points considered as noise)

3. KERNEL DESIGN ON BAGS OF BAGS

3.1 Framework

Let us denote $\mathbb{T}$ the set of all the video tubes of a film. A tube T_i is composed of chains C_{ri}. Each chain C_{ri} is a bag of SIFT vectors. Using set formulation:

$$T_i=\{C_{1i},\ldots,C_{ki}\} \text{ and } C_{ri}=\{SIFT_{1ri},\ldots,SIFT_{pri}\}$$

We want to design a kernel function $K(T_i,T_j)$ which will represent the similarity between two tubes. Proven it exists an embedding function $\Phi: \mathbb{T} \rightarrow \mathcal{H}$, which maps any bag T_i to a vector $\Phi(T_i)$ in a Hilbert space $\mathcal{H}$, one can define the kernel on bags K by a dot product in the induced space:

$$K(T_i,T_j) = <\Phi(T_i),\Phi(T_j)> \quad (1)$$

We consider, in this article, the general class of kernel on bags [5]:

$$K(T_i,T_j) = \sum_r \sum_s k(C_{ri},C_{sj}) \quad (2)$$

where k is the minor kernel as regard of major kernel on bags K. In our context, k is also a kernel on bags C_{ri} (of SIFT points). As proved in Chap. 9 of [5], such a function K is a kernel function as soon as k is a kernel function.

3.2 Minor Kernel on Bags

We have first to design the kernel function k between bags of SIFTs:

$$k(C_{ri}, C_{sj}) = <\phi(C_{ri}), \phi(C_{sj})> \quad (3)$$

with ϕ as embedding function of bags of SIFT into the feature space $\mathcal{H}$.

A simple solution could be to consider the linear kernel on bags of SIFT points:

$$k(C_{ri},C_{sj}) = \sum_{SIFT\in C_{ri}} \sum_{SIFT\in C_{sj}} <SIFT_{mri},SIFT_{nsj}> \quad (4)$$

The embedding function would then be explicitly defined by:

$$\phi(C_{ri}) = \sum_{SIFT\in C_{ri}} SIFT_{mri} \quad (5)$$

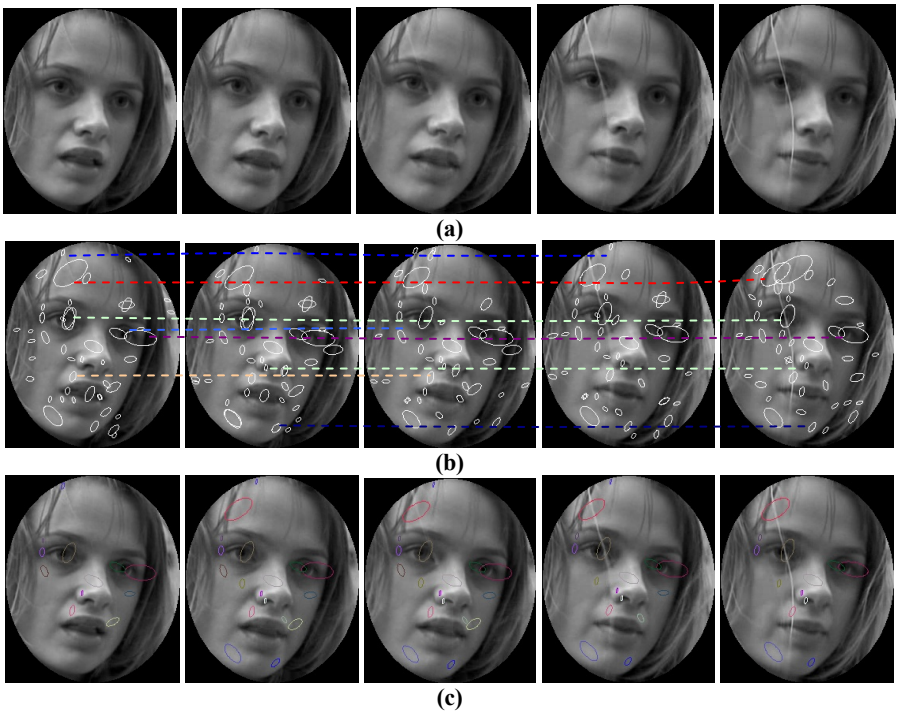

Figure 3 - An example of the tracking of five consecutive frames of faces
(a) original images (b) SIFT points filtering by temporal coherency (c) bags (of SIFT points) in different colours

However, in most cases we found (not extreme motion), all the feature points of the same chain are quite similar in the feature space, and consistent in scale and position. Thus, as embedding function of Eq.(5), we propose the averaging mapping function:

$$\phi(C_{ri}) = \frac{1}{|C_{ri}|} \sum_{SIFT \in C_{ri}} SIFT_{mri} \qquad (6)$$

where $|C_{ri}|$ represents the size (number of frames) of the chain C_{ri}.

Furthermore, instead of the linear kernel of Eq.(4), we combine the non-linear Gaussian χ^2 kernel, which has already proved its efficiency for image retrieval in [4], with the embedding function ϕ defined in Eq.(6) in order to design the minor kernel on bags:

$$k(C_{ri}, C_{sj}) = \exp\left(-\frac{1}{2\sigma^2} \frac{\left(\phi(C_{ri}) - \phi(C_{sj})\right)^2}{\phi(C_{ri}) + \phi(C_{sj})} \right) \qquad (7)$$

This function is still a kernel function as demonstrated in Chap.3 of [18].

Furthermore, our minor kernel is computed very fast, since we evaluate the average mapping function ϕ, defined in Eq.(6), on all the bags of SIFT from all the tubes, only once for all the retrieval process.

3.3 Major Kernel on Bags of Bags (Tubes)

The minor kernel k we obtained, can be replaced in the major kernel on tubes expression of Eq.(2). However, such an approach reduces the role of well-tracked SIFT points inside a tube, while these points are supposed to be more relevant since more persistent.

In order to take into account the size of the bag of feature vectors C_{ri} relatively to the tube size, we define the following weights:

$$w_{ri} = \frac{|C_{ri}|}{|T_i|} \qquad (8)$$

We exploit these weights in the formulation of our major kernel on bags of bags in Eq.(2):

$$K(T_i, T_j) = \sum_r \sum_s w_{ri} w_{sj} k(C_{ri}, C_{sj}) \qquad (9)$$

This key function has been proven to be still a kernel Gosselin et al. in [4].

Thanks to this new kernel scheme, the importance of well-tracked points is directly involved in the evaluation of similarity between tubes.

We have introduced our weighted minor kernel on bags into both "*lyu kernel*" on bags [8]:

$$K_{lyu}(T_i, T_j) = \frac{1}{|T_i|} \frac{1}{|T_j|} \sum_r \sum_s \frac{|C_{ri}|}{|T_i|} \frac{|C_{sj}|}{|T_j|} k(C_{ri}, C_{sj})^q \qquad (10)$$

and *Gosselin "pow-kernel"* on bags [4]:

$$K_{pow}(T_i, T_j) = \left(\sum_r \sum_s \frac{|C_{ri}|}{|T_i|} \frac{|C_{sj}|}{|T_j|} k(C_{ri}, C_{sj})^q \right)^{\frac{1}{q}} \qquad (11)$$

4. EXPERIMENTS

We have tried our actor retrieval framework, first, on small parts of the tv-serie "*Prison Break*" (from internet), then on the film DVD "*L'esquive*" of Abdellatif Kechiche (2004). We did our own ground truth on both videos. Unfortunately, at the current time, there is no real database of reference, publically available, on which we could make our tests and compare with other methods. The Visual Geometry Group of Oxford provides data, on which Sivic et al. [2] evaluated their approach, but these are only partial and pre-processed data, thus they are not relevant for us.

4.1 Visual Feature Extraction

We extracted 26 tubes of faces with 5 actors from "*Prison Break*" trailers. We tested our actor retrieval system on these tubes (Fig. 4), but the number of tubes for each actor was too small. Thus, we extracted 200 tubes of actor faces with 11 actors from the film "*L'esquive*". The mean number of faces in a tube is 54. The mean number of SIFTs in a tube is 169. We have put the 200 tubes, each represented by a bag of bags of SIFT vectors, into the interactive machine learning system of RETIN [19], with SVM core using our weighted kernels on bags (of bags). The variations of contexts are real ones as shown in the results (Fig. 5).

4.2 Performance Evaluation

We focus on retrieval results for a small learning set (less than 8 examples) to test the performance of our weighting on kernels on bags: kernel "*pow"* of Eq.(11) with q = 2 (Fig. 6) and kernel "*lyu"* of Eq.(10) with q = 2 (Fig. 7). We use the Gaussian χ^2 minor kernel in Eq.(7) with $\sigma = 1$ in each feature axe on normalized data. We can see on both curves that the kernels with our weight $\boldsymbol{w}$ are always more efficient than that without weight and this even with very few examples (2 examples). With both weighted kernels, MAP is more than 60% with only 2 tubes to learn the actor face. The weighted kernels are 5% better than classic ones with few examples.

We have tried our algorithm with different numbers of training set to obtain the curve of MAP (Mean Average Precision) of actor face retrieval. The MAP for kernel *pow* Eq.(11) is given on Fig. 8. We can see that the percentage of actor tube retrieved increases rapidly with the number of examples to reach a precision of more than 90% with 20 examples.

Furthermore, the weighted kernel, we propose here, is very fast to compute as explained in section 3.2.

5. CONCLUSION

In this paper, we have presented the design of new dedicated kernel on "bags of bags" for actor retrieval in films. From a film segmented into shots, we extracted face video tubes of actors and described these video objects with temporally coherent visual features. These features are then input into the RETIN system which provides an online machine learning framework. Our approach has been tested to retrieve actors in a real film and shown good results. The Mean Average Precision reaches more than 60% with only two exemplars of face tubes to learn the actor and reaches a precision of more than 90% with a small training set.

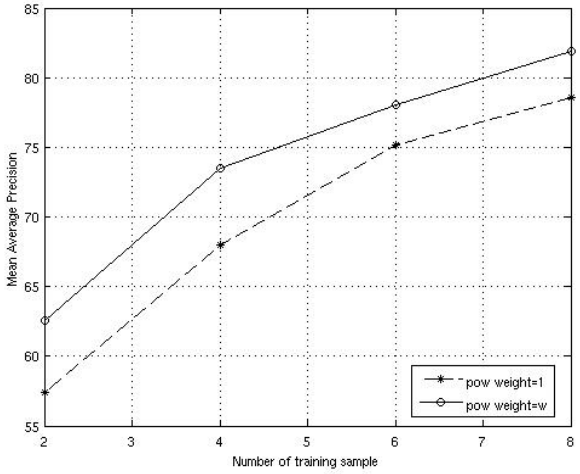

Figure 6 - MAP(%) of kernel on bags "*pow*" with/without weight for 2 to 8 training examples

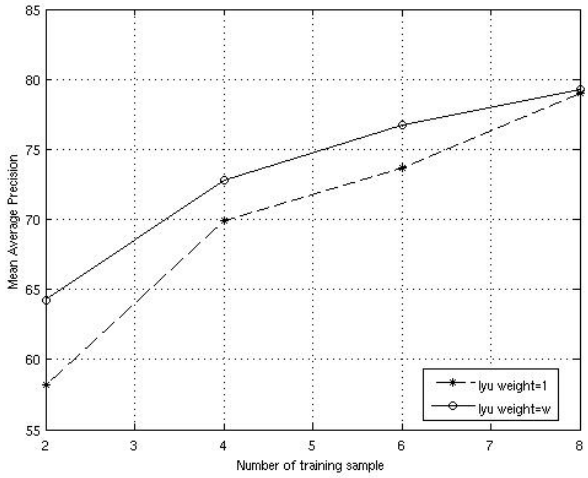

Figure 7 - MAP(%) of "*lyu*" kernel on bags with/without weight for 2 to 8 training examples

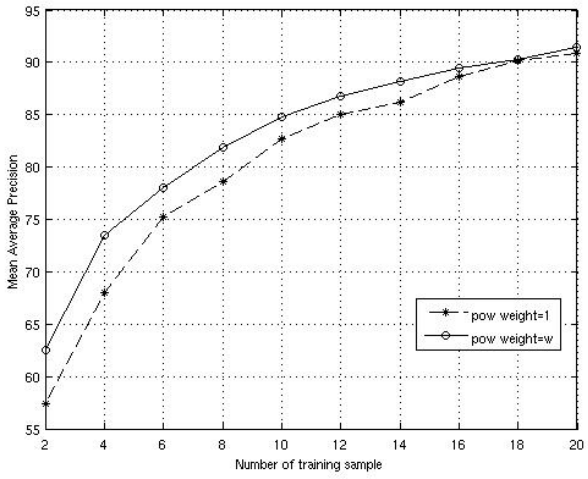

Figure 8 - MAP(%) of kernel on bags "*pow*" with/without weight for 2 to 20 training examples

One of the main issues we have to overcome is the lack of database. Indeed, as explained in the introduction, the few vi-

deo databases available are targeting security and biometrics applications. We are currently building a larger database with its ground-truth that we will make publically available for the community.

Next plans are to extract, from video-tubes, both global (colour, texture) and local visual features, in order to improve the robustness of our system. With such new features, we will have to design new kernel functions on bags of global and local features in the same framework.

6. ACKNOWLEDGMENTS

We want to thank Philippe-Henri Gosselin for providing the codes of kernels on bags within the retrieval system RETIN and for helping us with the experiments.

REFERENCES

[1] O. Arandjelovic, A. Zisserman, "Automatic Face Recognition for Film Character Retrieval in Feature-Length Films", *Proc. of IEEE CIVR*, 2005.

[2] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: video shot retrieval for face sets", *Proc. of IEEE CIVR,* 2005.

[3] R. Kondor and T. Jebara. "A kernel between sets of vectors", In *International Conference on Machine Learning (ICML)*, 2003

[4] P.H. Gosselin, M. Cord and S. Philipp-Foliguet. "Kernel on Bags for multi-object database retrieval", In *ACM International Conference on Image and Video Retrieval*, Amsterdam, The Netherlands, July 2007.

[5] J. Shawe-Taylor and N. Cristianini. "Kernel methods for Pattern Analysis", *Cambridge University Press*, ISBN 0-521-81397-2, 2004.

[6] C. Wallraven, B. Caputo, and A. Graf. "Recognition with local features: the kernel recipe", In *International Conference on Computer Vision (ICCV)*, volume 2, pages 257–264, 2003.

[7] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", *CVPR* 2001.

[8] S. Lyu. "Mercer kernels for object recognition with local features", In *IEEE CVPR*, San Diego, CA, 2005.

[9] D. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints", *IJCV*, 2004.

[10] M H Yang, D Kriegman, N Ahuja. "Detecting faces in images: A survey", *IEEE Trans on PAMI*, 2002, 24(1):34-58.

[11] E. Casiraghi, Raffaella Lanzarotti, Giuseppe Lipori, "A face detection system based on color and support vector machines", *Springer Berlin / Heidelberg*, Volume 2859, 2003 pages 113-120.

[12] Chen, J. Shan, S. Yang, P. Yan, S. Chen, X. Gao, W., "Novel Face Detection Method Based on Gabor Features", *Springer Berlin / Heidelberg*, 2004, ISSU 3338, pages 90-99.

[13] R. Lienhart, A. Kuranov, V. Pisarevsky, "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection", *MRL Technical Report*, May 2002.

[14] *http://www.face-rec.org/databases/*

[15] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, "Face Recognition: A Literature Survey", *ACM Computing Surveys*, 2003, pp. 399-458.

[16] G. Camara Chavez, F. Precioso, M. Cord, S. Philipp-Foliguet, A. de Albuquerque Araujo, "Shot boundary detection by a hierarchical supervised approach", *IEEE IWSSIP*, 2007.

[17] P.H. Gosselin, M. Cord and S. Philipp-Foliguet. "Kernel on Bags of Fuzzy Regions for fast object retrieval", *In IEEE ICIP,* San Antonio, Texas, USA, September 2007.

[18] N. Cristianini and J. Shawe-Taylor. "An introduction to Support Vector Machines and other kernel-based learning methods", *Cambridge University Press*, ISBN 0-521-78019-5, 2000.

[19] P.H. Gosselin, M. Cord, S. Philipp-Foliguet, "Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval", *CVIU, in press,* 2008.

[20] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study", *International Journal of Computer Vision*, Vol.73, No. 2, pp. 213-238 2007.

Figure 4 - Examples of actor face tubes extracted from the film "*Prison Break*" in our actor retrieval system based on RETIN system, one image represents one tube.

(a) (b) (c) (d)

Figure 5 - Results of our interactive actor retrieval system based on RETIN system, for the film "*L'esquive*", one image represents one tube : (a) Query initialization with one request; (b) First results: tubes ranked regarding the tubes similarities; (c) Second iteration with one more positive example(green squares■) and two negative examples (red squares■); (d) Results after 2 iterations.