SPATIO-TEMPORAL TUBE KERNEL FOR ACTOR RETRIEVAL*

Shuji Zhao, Frédéric Precioso

ETIS, CNRS, ENSEA
Univ Cergy-Pontoise, France
zhao,precioso@ensea.fr

Matthieu Cord

LIP6, CNRS
UPMC, France
matthieu.cord@lip6.fr

ABSTRACT

This paper presents an actor video retrieval system based on face video-tubes extraction and representation with sets of temporally coherent features. Visual features, SIFT points, are tracked along a video shot, resulting in sets of feature point chains (spatio-temporal tubes). These tubes are then classified and retrieved using a kernel-based SVM learning framework for actor retrieval in a movie. In this paper, we present optimized feature tubes, we extend our feature representation with spatial location of SIFT points and we describe the new Spatio-Temporal Tube Kernel (*STTK*) of our content-based retrieval system. Our approach has been tested on a real movie and proved to be faster and more robust for actor retrieval task.

Index Terms— Face recognition, Video object, Actor retrieval, Kernel on bags, Spatio-Temporal Tube Kernel.

1. INTRODUCTION

Content-Based Image Retrieval (CBIR) has attracted a lot of research interest in recent years. Significant progress has been achieved in the performance of object categorization and retrieval systems in the domain of multimedia retrieval. In this paper we focus on video and retrieval of actors in movies.

Recent actor retrieval systems are following quite similar processing chain:

- Face detection: this step is based on Viola & Jones detector [1], or derived versions, in most works.
- Feature extraction: face segmentation resulting in a face video tube then local visual feature extraction from this video tube.
- Feature representation: one feature vector, a set of vectors or more complex spatial structures to represent video tubes.
- Classification: methods based on projections onto a visual dictionary or based on machine learning techniques.

*THIS WORK IS FUNDED BY *K-VIDEOSCAN* DIGITEO PROJECT NO. 2007-3HD AND *ITOWNS* ANR MDCO 2007 PROJECT.

Fig.1 shows our actor retrieval framework. Instead of considering precise facial features (eye, mouth, etc.) as in related works [2][3], we want to avoid introducing prior knowledge and thus focus on more generic feature which make our system adaptable to other semantic video objects, eg. cars. Furthermore, kernel-based methods allow us to exploit recent machine learning techniques as active learning.

In [4] a video object is represented by a set of temporally consistent chains of local descriptors SIFT (a bag of bags of features) without considering the position of each chain. Integration of spatial information became recently more important in CBIR research works as illustrated by Fergus et al. [5] *constellation model*, Felzenszwalb et al. [6] *pictorial model* or Heisele et al. [7] *object parts*. In kernel-based framework also propositions have been made with Pyramid kernel [8] or kernel on pairs of regions [9]. In this paper, we design a new kernel in order to integrate spatial information in our spatio-temporal tubes and provide optimizations to tube extraction.

2. TUBE EXTRACTION OPTIMIZATION

In [4] the faces of actors are detected by Viola & Jones detector [1] and segmented by ellipses which approximate face contours. A video object is then defined by a video tube made of face regions in the successive frames of a shot. From a face video tube, we extract a set of temporally consistent chains of local descriptors SIFT, that we call a spatio-temporal (feature) tube.

One of the main issues with considering such feature tube representation lies in the size of data to process. In this paper, we propose three improvements to enrich while reducing our representation of the visual features: (a) parameter optimization of descriptor SIFT to make it adaptive to the data; (b) intra-tube chains tracking to obtain more consistent and more compact chains for each video tube and thus reduce computational complexity; (c) tubes with Spatial Position.

2.1. Adaptive parameter of SIFT descriptor

For the extraction of descriptor SIFT, we found out that the number of SIFT points extracted is quite sensitive to the scale of the "first octave" (see [10]). We optimize the scale of the

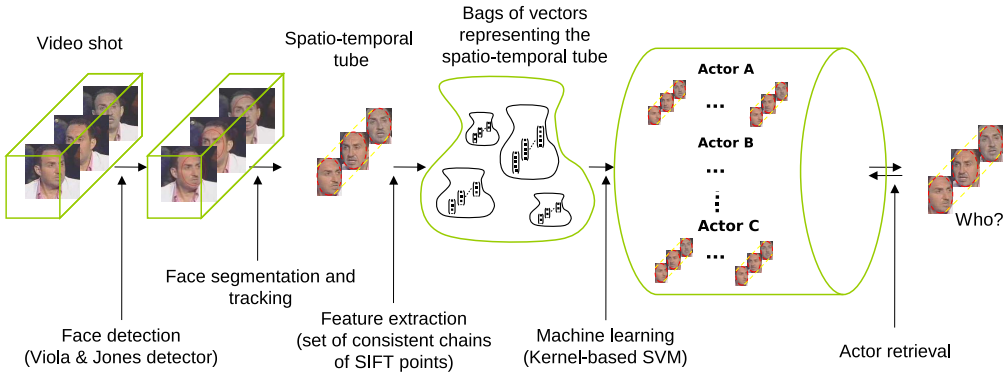


Fig. 1. STTK-based actor retrieval system.

first octave by multiplying the scale of face image by a coefficient $\lambda = 2^n (n = \ldots -2, -1, 0, 1, 2 \ldots)$. n is selected to set the scale of the first octave in a certain interval (50 to 100 pixels in width), hence we can extract SIFT points even if the image is small and reduce the number of irrelevant points extracted in big images.

2.2. Intra-Tube Chain Tracking

In order to improve the consistency of chains and reducing the numbers of chains in a tube, we propose to link together several short chains into one long chain (Fig.2). The intra-

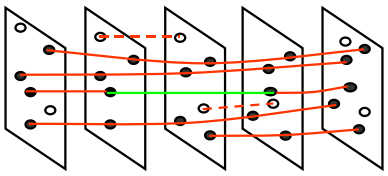


Fig. 2. Intra-tube chain tracking (solid lines: consistent chains, dash lines: noise, green lines: link of two short chains)

tube tracking is achieved by matching two short chains if their average SIFT vectors are similar enough (L^2 distance below 200 in our case) and average normalized positions of the two chains close enough (below 0.2 in our work). See Fig.4 for the definition of normalized positions. Thus, the chains of descriptors SIFT become more consistent and the number of chains per tube is highly reduced. See Fig.3 (a)(b) for two examples of intra-tube tracking. To evaluate the consistency of the SIFT descriptor along a chain, we show temporal stability of some long chains of SIFT with the images of Fig.3 (c)(d)(e). One line in any of these images represent a 128-dimensional SIFT vector while in column you can see the variation of one of these 128 values along its tracked chain.

2.3. Tubes with Spatial Position

We introduce the position of each SIFT chain in the representation of the tube, so that we strengthen the comparison between same parts of face. The position of a chain is defined by the mean normalized position of SIFT points (x, y) in the chain, see Fig. 4.

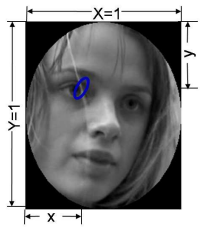


Fig. 4. Normalized position of SIFT point. Scale and orientation of the ellipse represent scale and orientation of the SIFT point.

3. SPATIO-TEMPORAL TUBE KERNEL (*STTK*)

3.1. Kernel on Tubes (Without Position)

Let us denote T_i a tube, C_{ri} a chain and $SIFT_{mri}$ a SIFT vectors. Using set formulation: $T_i = \{C_{1i}, \ldots, C_{ki}\}$ and $C_{ri} = \{SIFT_{1ri}, \ldots, SIFT_{pri}\}$, we have designed our major kernel on tubes in [4] by weighted *"power-kernel"* [9]:

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

where $|C_{ri}|$ represents the size (number of frames) of the chain C_{ri}; $|T_i|$ represents the size of the tube T_i; $k(C_{ri}, C_{sj})$ is the minor kernel on chains, a Gaussian χ^2 kernel:

$$k(C_{ri}, C_{sj}) = exp\left(-\frac{1}{2\sigma_1^2} \frac{\left(\overline{C}_{ri} - \overline{C}_{sj}\right)^2}{\overline{C}_{ri} + \overline{C}_{sj}} \right) \quad (2)$$

3.2. Spatio-Temporal Tube Kernel (*STTK*)

The kernel of Eq.(1) defines a similarity function considering exhaustively all the chains of tube T_i with all the chains of

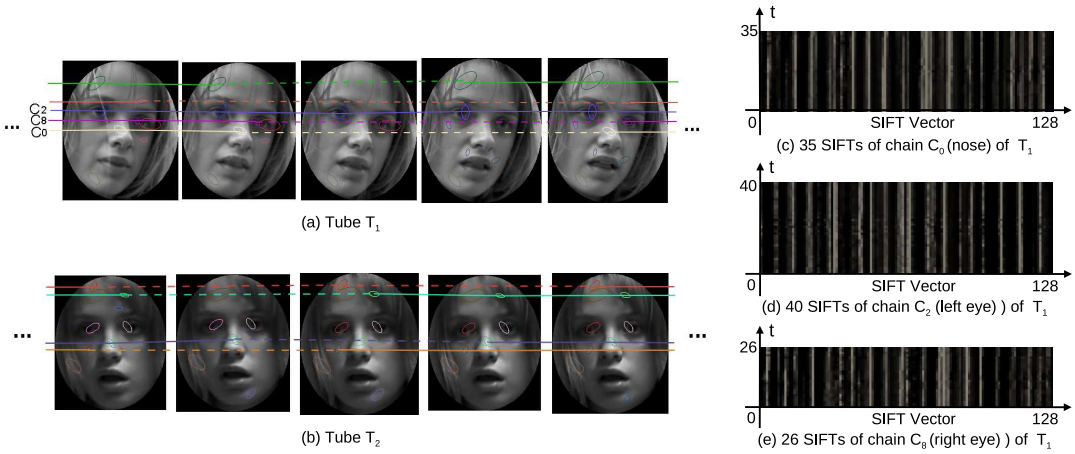


Fig. 3. Intra-tube chains tracking. (a)(b) Example of two tubes, SIFT points along the same chain are in same color (the scale and orientation of ellipses represent the scale and orientation of SIFT). (c)(d)(e) Consistency of three SIFT chains of tube T_1.

tube T_j, wherever are their positions on the face. That is to say, we compare not only the left eye chain of tube T_i and left eye chain of tube T_j, but also left eye chain of tube T_i and mouth chain of tube T_j. In this paper, we redefine the minor kernel on chains by adding a term taking into account the relative positions of two chains:

$$k'(C_{ri}, C_{sj}) = k(C_{ri}, C_{sj}) e^{-\frac{(\overline{x}_{ri}-\overline{x}_{sj})^2+(\overline{y}_{ri}-\overline{y}_{sj})^2}{2\sigma_2^2}} \quad (3)$$

where $k(C_{ri}, C_{sj})$ is the previous minor kernel in Eq.(1); $(\overline{x}_{ri}, \overline{y}_{ri})$ is the mean position of SIFT points in chain C_{ri} of tube T_i.

Furthermore, we propose to improve the importance of lengthy chains, as well as lengthy tubes, by modifying the weights on chains: $|C_{ri}|/|T_i|$ into $|C_{ri}|/\sqrt{|T_i|}$.

Hence, the major kernel on tubes becomes:

$$K'_{pow}(T_i, T_j) = \left(\sum_r \sum_s \frac{|C_{ri}|}{\sqrt{|T_i|}} \frac{|C_{sj}|}{\sqrt{|T_j|}} k'(C_{ri}, C_{sj})^q \right)^{\frac{1}{q}} \quad (4)$$

With this new kernel on tubes, we improve the importance of the comparison between two chains approximately at the same position, eg. left eye chain of tube T_i and left eye chain of tube T_j. For the comparison between two chains at much different positions, eg. left eye chain of tube T_i and mouth chain of tube T_j, the weight is reduced. Thus, the importance of this matching in the evaluation of the similarity is also lowered.

4. EXPERIMENTS

We have tested our actor retrieval framework on the same database of [4], 200 faces tubes of 11 actors from the movie *"L'esquive"*. The mean number of faces in a tube is 54. These 200 tubes have been used as input to the interactive machine learning system of RETIN [11], with SVM core using our weighted kernel on tubes. The interactive retrieval process with RETIN is presented in Fig.6. We have evaluated the retrieval precision for 3 actors from the database and proved the efficiency both in reducing the calculate time and in improvement of precision comparing with the work of [4].

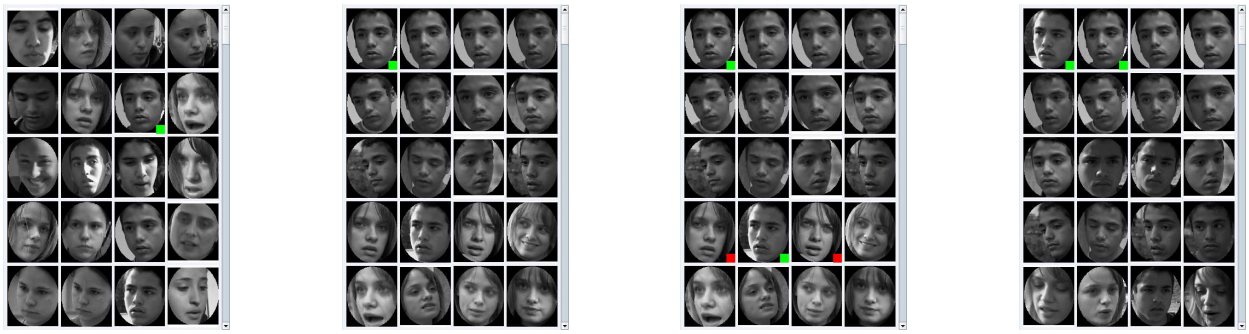
4.1. Optimized Feature Extraction

From each of the 200 tubes, we have extracted the visual features step by step: (a) Extraction of descriptors SIFT for each image. We have optimized the parameters of SIFT to reduce the number of chains for each tube. The mean number of SIFT chains in a tube is reduced from an average of 169 (result of [4]) to an average of 64; (b) Intra-tube chain tracking. This enable us to get more consistent and more compact chains of descriptors SIFT. The average number of SIFT chains in a tube is then reduced again from 64 to 49.

For comparing a tube of n chains with another tube of m chains, the number of comparisons in our kernel is $n \times m$ (see Eq.4). Thus, with (a) and (b), the new system is about $\left(\frac{169}{49}\right)^2 \approx 12$ times faster than our previous system. We have evaluated the impact of these modification on visual feature extraction. The results are, if not identical as previous ones, even a little better thanks to the reduction of noise by removing small chains, see Fig.5 for the comparison of MAP(Mean Average Precision) for one actor.

4.2. Spatio-Temporal Tube Kernel (*STTK*) Evaluation

We have extracted not only tubes of SIFT vectors, but also the position of each SIFT location in face image. We then input these spatio-temporal tubes of SIFT vectors integrating spatial position, into the interactive kernel-based retrieval system RETIN [11], with our *STTK* kernel function (Eq.4). We have also tried to compare the distance between the positions of



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

Fig. 6. 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.

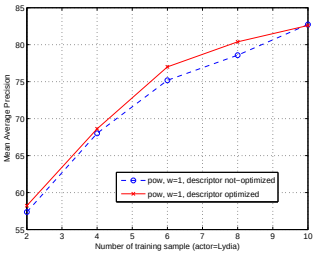
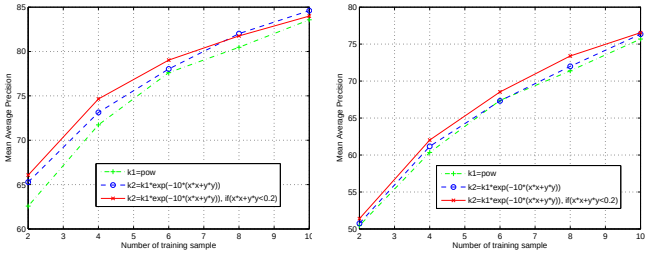


Fig. 5. MAP(%) of kernel on tubes K_{pow} for short / long chains, $q = 2, \sigma_1 = 1$.

two chains before computing the kernel. We compare only pairs of chains which are not too far ($d^2 < 0.2$ in our experiments). We thus reduce again the complexity of our algorithm.

We focus on retrieval results for a small training set (less than 10 examples). The MAP curves of retrieval of one actor (Lydia), see Fig.7(a), and the MAP curves of retrieval averaged over three actors (Lydia, Crimon and Hanane), see Fig.7(b), show that our *STTK* method is more efficient than a kernel on spatio-temporal tubes without integrating spatial information.



(a) (b)

Fig. 7. MAP(%) of kernel on tubes K_{pow} (without position) and K'_{pow} (with position), $q = 2, \sigma_1 = 1, \sigma_2 = \sqrt{0.05}$. (a) result for one actor (b) result averaged on 3 different actors

5. CONCLUSIONS

In this paper, we have presented a efficient actor retrieval system, which considered a face video tube as a video object. From a video tube we extracted a "tubes" of visual features as well as spatial location of features. The design of a new kernel embedding this spatial constraint has been proved to be more powerful for actor retrieval in a real movie. Next step will be to devise kernel functions, like Fischer kernels, from generative models as *constellation model* or *pictorial model* mentioned in the introduction.

6. REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001, pp. 511–518.

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

[3] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" – automatic naming of characters in TV video," in *BMVC*, 2006.

[4] S. Zhao, F. Precioso, M. Cord, and S. Philipp-Foliguet, "Actor retrieval system based on kernels on bags of bags," in *EUSIPCO*, Lausanne, Switzerland, 2008.

[5] R. Fergus, P. Perona, and A. Zisserman, "A sparse object category model for efficient learning and exhaustive recognition," in *CVPR*, 2005.

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008.

[7] B. Heisele, T. Serre, M. Pontil, and T. Poggio, "Categorization by learning and combining object parts," in *NIPS*, Vancouver, 2001.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[9] P.-H. Gosselin, M. Cord, and S. Philipp-Foliguet, "Kernel on bags for multi-object database retrieval.," in *CIVR*, Amsterdam, The Netherlands, July 2007.

[10] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *IJCV*, 2003, vol. 20, pp. 91–110.

[11] P.-H. Gosselin and M. Cord, "Active learning methods for interactive image retrieval.," *IEEE Trans. on Image Processing*, vol. 17, no. 7, pp. 1200–1211, 2008.