論文 / 著書情報
Article / Book Information

Title	High-Level Feature Extraction Using SIFT GMMs and Audio Models
Author	Nakamasa Inoue, Tatsuhiko Saito, Koichi Shinoda, Sadaoki Furui
Journal/Book name	Proc. ICPR2010, Vol. , No. , pp. 3220-3223
発行日 / Issue date	2010, 8
権利情報 / Copyright	(c)2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

2010 International Conference on Pattern Recognition

High-Level Feature Extraction Using SIFT GMMs and Audio Models

Nakamasa Inoue, Tatsuhiko Saito, Koichi Shinoda, and Sadaoki Furui
Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552, Japan
{inoue,saito}@ks.cs.titech.ac.jp, {shinoda,furui}@cs.titech.ac.jp

***Abstract*—We propose a statistical framework for high-level feature extraction that uses SIFT Gaussian mixture models (GMMs) and audio models. SIFT features were extracted from all the image frames and modeled by a GMM. In addition, we used mel-frequency cepstral coefficients and ergodic hidden Markov models to detect high-level features in audio streams. The best result obtained by using SIFT GMMs in terms of mean average precision on the TRECVID 2009 corpus was 0.150 and was improved to 0.164 by using audio information.**

I. Introduction

Recently, a large amount of video data has become available on the Internet. Finding interesting or necessary parts in a video requires an automatic technique based on statistical pattern recognition for video indexing. Considering that high-level feature (HLF) extraction is an important topic in computer vision, our objective in this work was to extract HLFs, which are human-recognizable objects, events and scenes. For example, "Airplane", "Boat Ship", "People dancing", "Singing", "Cityscape", and "Nighttime" were chosen as target HLFs in the 2009 TRECVID workshop [1].

The bag-of-visual-words (BoW) approach is one of the most successful methods for attaining general object recognition and is also useful for HLF extraction [2]. In this method, local features such as SIFT features [3] are extracted from an image and quantized as visual words by applying clustering techniques. Then a BoW histogram is made by counting appearances of visual words. The soft-assignment BoW approach, which can estimate distribution precisely, has been receiving particular attention recently [4], [5], [6]. This process requires a large number of local features; however, most video-based HLF extraction methods extract features from only key-frame images. Improving accuracy requires a multi-frame approach, which is expected to work well when images of objects are taken from different angles.

Combinations of visual and audio schemes have been proposed in advanced researches [7], [8]. Jiang *et al.* [7] use a matching pursuit method and make a joint audio-visual codebook. Snoek *et al.* [8] propose an audio segmentation module that includes several components providing such functions as acoustic change detection and speech/non-speech classification. Both visual and audio information are important for extracting certain HLFs, such as "Singing", "People playing a musical instrument" and "People dancing".

We propose a combination of SIFT GMMs and audio models as an effective means for HLF extraction. In our approach, we extract SIFT features from not only key frames but all image frames in a shot. Then, we model the features by a GMM in order to estimate distribution precisely. In the audio part, we extract mel-frequency cepstral coefficients (MFCCs) and model HLFs by ergodic hidden Markov models (HMMs). MFCCs are short-term spectral features that are widely used in the field of speech recognition. Furthermore, the effectiveness of MFCCs for audio classification has been reported in previous researches. Since there are HLFs that can be easily detected in audio streams, MFCCs are expected to be effective for HLF extraction.

This paper is organized as follows. Section II describes visual models that use SIFT GMMs. Section III introduces audio models and a combination of the two schemes. Section IV evaluates our system on the TRECVID 2009 data set. Section V concludes the paper with a brief summary of key points and mentions future work to be done.

II. SIFT Gaussian Mixture Models

A. Visual feature extraction

SIFT [3] is a local feature extraction algorithm that is widely used for general object recognition. The SIFT features are invariant to image scale and robust against changes in illumination and noise.

The feature extraction steps are divided into two parts: local region extraction and feature description. For the first part, many extraction methods have been proposed such as DoG, Harris-Laplace, dense, and random. Our approach is to extract Harris-Affine regions [9] and Hessian-Affine regions [10]. Since they are invariant to affine transformations, they are expected to be robust against camera angle changes. In the second part, SIFT features are extracted from each region. We apply PCA to reduce the dimension, using 32 dimensional SIFT features where the original feature dimension is 128. We extract SIFT features from not only key frames but all image frames in a shot.

B. SIFT Gaussian mixture models

We model SIFT features extracted from each shot by using a GMM, referring to the resulting GMM as a SIFT GMM. The probability density function (pdf) of a SIFT GMM is

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.787

given by

$$p(x|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k), \quad (1)$$

where K is the number of mixtures, $\theta = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ is a set of parameters, including a mixing coefficient w_k and a pdf of Gaussian distribution $\mathcal{N}(x|\mu_k, \Sigma_k)$ with mean vector μ_k and variance matrix Σ_k.

The parameters of a GMM are often estimated by using the expectation maximization (EM) algorithm. The EM algorithm is known as a method for finding maximum likelihood estimators of a model with latent variables. However, the number of SIFT features in each shot may not be enough to estimate parameters precisely. Thus, we estimate the mean vector by using the maximum a posteriori (MAP) adaptation technique. A universal background model (UBM) estimated using all the video data is used to obtain the a priori distribution. We use a Gaussian distribution for the a priori distribution. Mean vectors are adapted by

$$\mu_k = \frac{\tau_k \mu_k^{(g)} + \sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\tau_k + \sum_{n=1}^{N} \gamma(z_{nk})} \quad (2)$$

where $\gamma(z_{nk})$ is the posterior probability of the Gaussian component k for x_n, $\mu_k^{(g)}$ is a mean vector of the UBM, and τ_k is a hyper-parameter.

C. Kernel SVM classification

In the classification part, we first compute the distance between shots, which is defined by the weighted sum of Mahalanobis distances between the corresponding mixture components. Since pairs of the corresponding Gaussian components are given in the MAP adaptation step, the distance between s-th and t-th shots is given by

$$d(s,t) = \sum_{k=1}^{K} w_k^{(g)} (\mu_k^{(s)} - \mu_k^{(t)})^T (\Sigma_k^{(g)})^{-1} (\mu_k^{(s)} - \mu_k^{(t)}), \quad (3)$$

where $\theta^{(g)} = \{w_k^{(g)}, \mu_k^{(g)}, \Sigma_k^{(g)}\}_{k=1}^{K}$ is a parameter set of the UBM and $\theta^{(s)}$ and $\theta^{(t)}$ are parameter sets of GMMs of the s-th and t-th shots, respectively.

Finally, the shots are classified by using support vector machines (SVMs). We use an RBF kernel given by

$$K(s,t) = \exp(-\gamma d(s,t)), \quad (4)$$

where γ is a parameter optimized through experiments. Then, the posterior probability $p(h = +1|X_s)$ is estimated by the SVM with probability outputs, where h is a random variable that is equal to $+1$ if the target HLF appears in the s-th shot.

We extract features with Harris-Affine regions and Hessian-Affine regions independently, denoting the resulting posterior probabilities as $p_{\text{har}}(h = +1|X_s)$ and $p_{\text{hes}}(h = +1|X_s)$, respectively.

III. Audio Models

A. Audio feature extraction

We extract MFCCs (12 dimensional) as audio features from 100 msec Hamming-windowed frames with 50 % overlap. We also use ΔMFCCs (12 dim) and ΔΔMFCCs (12 dim), which are first and second order derivatives of MFCCs, respectively, and Δlog-power (1 dim) and ΔΔlog-power (1 dim). The total dimension of the features is 38.

B. Hidden Markov models

We model each HLF using an ergodic HMM that has a Gaussian distribution for each state and transition probabilities between all states. An ergodic HMM is made for each target HLF and a UBM is estimated by using shots that do not include any target HLFs. We use the EM algorithm to estimate the HMM parameters.

In the detection part, we use a likelihood ratio between the target HMM and the UBM:

$$l_{\text{au}} = \frac{p_{\text{au}}(X_s|h = +1)}{p_{\text{au}}(X_s|h = -1)}, \quad (5)$$

where $p_{\text{au}}(X_s|h = +1)$ is a likelihood from the target HMM and $p_{\text{au}}(X_s|h = -1)$ is a likelihood from the UBM.

C. Combination of SIFT GMMs and audio models

We combine SIFT GMMs and audio models by a combined log likelihood ratio L given by

$$L = w_{\text{au}} L_{\text{au}} + w_{\text{har}} L_{\text{har}} + w_{\text{hes}} L_{\text{hes}}, \quad (6)$$

where L_{au} is a log likelihood ratio from audio models and L_{har} and L_{hes} are log likelihood ratios from SIFT GMMs with Harris-Affine regions and Hessian-Affine regions, respectively. $w_{\text{au}}, w_{\text{har}}$ and w_{hes} are weights for each stream.

In order to compute L_{har} from the posterior probability $p_{\text{har}}(h = +1|X_s)$, we apply Bayes' theorem as follows.

$$\begin{aligned} L_{\text{har}} &= \log \frac{p_{\text{har}}(X_s|h = +1)}{p_{\text{har}}(X_s|h = -1)} \\ &= \log \frac{p_{\text{har}}(h = +1|X_s)}{p_{\text{har}}(h = -1|X_s)} \cdot \frac{p_{\text{har}}(h = -1)}{p_{\text{har}}(h = +1)} \\ &= \log \frac{p_{\text{har}}(h = +1|X_s)}{1 - p_{\text{har}}(h = +1|X_s)} + const. \end{aligned} \quad (7)$$

The first term in the right-hand side of Eq. (7) is a log odds ratio. L_{hes} is computed in the same way.

IV. Experiments

A. Experimental conditions

Our experiments were conducted using the TRECVID 2009 development video data set [2]. The set mainly comprises documentaries and educational programs developed by the Netherlands Institute for Sound and Vision. We used roughly half (18,120) the shots in them for training and the rest (18,142) for testing. There were 20 types of target HLFs

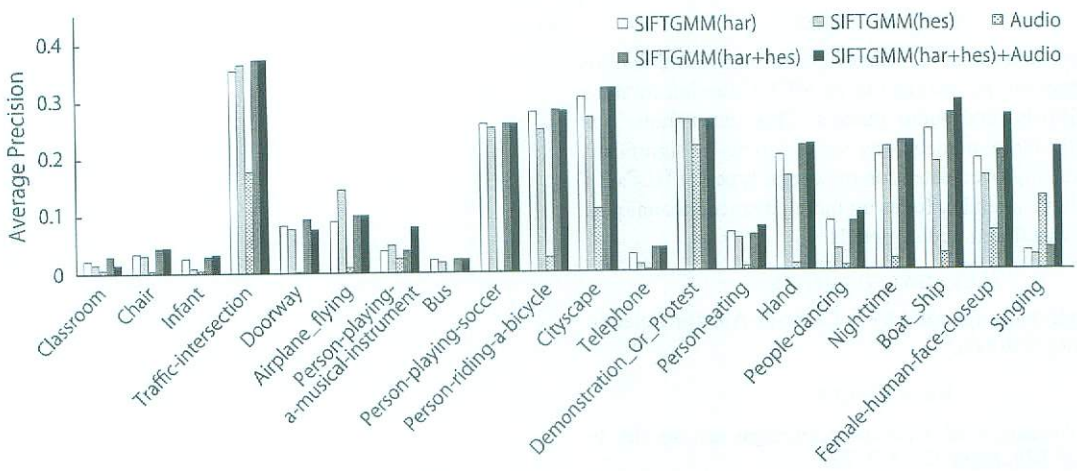

Figure 1. Detection performance in terms of average precision (AP) on the TRECVID 2009 data set. "SIFT GMM (har)" and "SIFT GMM (hes)" indicate results obtained by using Harris-Affine and Hessian-Affine region detectors, respectively. "Audio" uses the audio models only. "SIFT GMM (har+hes)" is the combination of the two region detectors. "SIFT GMM (har+hes)+Audio" is the fusion of all.

Table I
MEAN APS FOR DIFFERENT SCHEMES.

	Mean AP
SIFT GMM (har)	0.141
SIFT GMM (hes)	0.129
Audio	0.042
SIFT GMM (har+hes)	0.150
SIFT GMM (har+hes)+Audio	0.164

in the TRECVID 2009; we used average precision (AP) and Mean AP (the average AP for all the HLFs) as measures for evaluating them.

The SIFT GMMs we used comprised 512 mixture components, which corresponds to the visual word vocabulary size. Each of the HMMs had two states and 512-mixture-component GMMs were used for each state. We used 2-fold cross-validation to optimize the weight parameters w_{au}, w_{har}, and w_{hes}.

B. Results

Figure 1 and Table I show detection performance for the TRECVID 2009 data set. Combining the two region detectors relatively improved the performance by 6.1% and the audio models improved it by a further 10.0%. Particularly significant improvement was obtained for the APs of "Singing", "Person playing a musical instrument", and "Female human face closeup" (458%, 106%, and 30.0%, respectively). The optimized weight of the audio (w_{au}) exceeded that of the visual only ($w_{\mathrm{har}}+w_{\mathrm{hes}}$) for these three HLFs. Since the female voice is pitched higher than that of males, the audio scheme helped reject male human faces. This shows that audio information is essential for some kinds of HLFs.

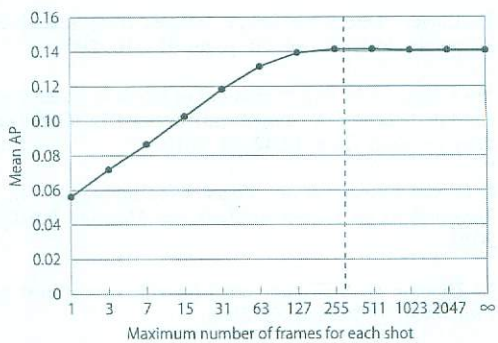

Figure 2. Comparison of Mean APs with different numbers of frames. The dotted line indicates the average number of frames in a shot.

We thinned some of the image frames as a means of reducing computational time. Figure 2 shows a comparison of Mean APs with different numbers of frames in each shot. The results were obtained by using SIFT GMMs with Harris-Affine regions. There are, on average, 270 frames in each shot and 8665 frames in the longest shot. We obtained nearly identical accuracies (~50 %) by using 127 frames and using all the frames.

The performance we obtained with our system ranked 4th among all the results presented by teams participating in TRECVID 2009. Significantly, our results for "Singing" and "People dancing" accuracy ranked 1st. While the other systems presented used short-term (~100 ms) representation for audio streams (e.g.[7]), ours used a model that represents long-term tendencies in video. This is the main reason we consider our system to be more advantageous than others in terms of HLF detection.

V. Conclusion

We proposed a statistical framework for high-level feature (HLF) extraction that makes use of SIFT Gaussian mixture models (GMMs) and audio models. This multi-frame approach with audio models was shown to be a significant step towards improved detection of several types of HLFs. In future work we intend to focus on more advanced techniques for fusing visual and audio models.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientic Research (B) 20300063.

References

[1] A. F. Smeaton, *et al.* Evaluation campaigns and trecvid. In *Proc. of MIR*, pages 321–330, 2006.

[2] A. F. Smeaton, *et al.* High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In *Multimedia Content Analysis, Theory and Applications*, pages 151–174, Springer Verlag, 2009.

[3] D. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, vol. 20, pages 91–110, 2004.

[4] Y.-G. Jiang, *et al.* Representations of keypoint-based semantic concept detection: A comprehensive study. In *IEEE Transactions on Multimedia*, 12:42–53, 2010.

[5] J. C. van Gemert, *et al.* Visual word ambiguity. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[6] X. Zhou, *et al.* SIFT-bag kernel for video event analysis. In *Proc. of ACM Multimedia*, 2008.

[7] W. Jiang, *et al.* Short-term audio-visual atoms for generic video concept classification. In *Proc. of ACM Multimedia*, 2009.

[8] C. G. M. Snoek, *et al.* The mediamill trecvid 2009 semantic video search engine. In *Proc. of TRECVID workshop*, 2009.

[9] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. of ECCV*, pages 128–142, 2002.

[10] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. In *IJCV*, vol. 60(1), pages 63–84, 2004.