A COARSE-TO-FINE LOGO RECOGNITION METHOD IN VIDEO STREAMS

Chaoyang Zhao, Jinqiao Wang, Chengli Xie, Hanqing Lu

National Laboratory of Pattern Recognition, CASIA, Beijing China, 100190
{chaoyang.zhao, jqwang, clxie, hqlu}@nlpr.ia.ac.cn

ABSTRACT

Visual logo recognition is significant for many applications, such as enterprise identification, entertainment advertising, vehicle recognition, road sign reading, trademark protection, and much more. In this paper, we propose a coarse-to-fine framework to recognize visual logos from video streams. To reduce the instability of the initial template selection problem, we introduce the "iconic template" selection strategy to select effective template set for visual logos. At the coarse stage, we adopt DOT(Dominant Orientation Templates) matching with a low threshold to find logo candidates. At the fine stage, we transform the multiple template matching problem into a pairwise binary classification problem. The candidates collected from the template matching process combined with the target template are send to a pairwise binary classifier to predict whether the candidate and the template belong to the same logo or not. The pairwise binary classifier is trained in an offline manner and with an unsupervised training data collection strategy. The proposed method can flexibly adapt to different template matching approaches and various matching thresholds. The false-alarm rate is greatly reduced through the second stage. Experimental results show the feasibility and effectiveness of the proposed approach.

Index Terms— logo recognition, logo detection, template matching, pairwise learning

1. INTRODUCTION

According to the United States Patent and Trademark Office, there are nearly 0.25 million new logo registrations in 2012 and increasing rapidly. As an active research area, logo recognition from video streams is vital for many commercial applications. It can be widely used for registered trademark protection to avoid a likelihood of confusion. Besides, it can also be applied to product monitor systems embedded in TV shows or online video websites for watching sport videos, online movies, etc. The duration and frequency of the logo appearing in broadcast are important factors to assess the effect of advertisements, which is directly related to adjustment of the commercial investment. The size, position and deformation of the logo are important to attract customers' attention and evaluate the performance of advertisement. Automatically detecting logos can significantly reduce the intensive workload of broadcast TV monitoring. Generally speaking, logos are composed of text, graphics and storytelling images. Usually, the text has the name of the company and/or information of the product. The graphics and images create a more or less abstract, symbolic, or vivid indicator of the product or the company.

In the past decades, many researches on logo recognition tended to follow the improvements of general object recognition approaches[1, 2, 3, 4] which provided promising results. Traditional logo detection often leveraged on object detection approaches, which used statistical learning to build a classifier offline and later used it at run-time for the recognition[4]. This works fine when discriminating one category from others(e.g. car, or pedestrian detection). However, when recognition needs to be efficient with high accuracy, for example, recognition of traffic sign logos for a driverless car, this approach usually suffers from limitations when running in a real-time application or confusion for different signs with small intra-class variability like similar shapes(e.g. turn left sign and turn right sign shown in Fig. 2).

Template matching is widely used for logo recognition because of its simplicity and capability to handle different types of objects. Many state-of-the-art approaches for template matching have been used for logo detection and recognition[2, 5, 3]. There are two problems limits the performance. One is the problem of initial template selection. The existing template matching approaches often manually select an initial template, which significantly varies the performance of logo recognition. The other one is how to select a proper threshold to reduce missed matches and false alarms. For learning based approaches, there are much fewer positive samples available since the appearances are usually similar. It lacks of significant differences for the same logo. As the analysis above, to combine the learning based approaches with template matching, we propose a coarse to fine framework in this paper. In particular, iconic template matching and pairwise learning methods are combined to recognize multiple logos in a coarse-to-fine manner. Instead of applying a matching threshold, a binary classification method is utilized to judge whether the detectd logo and the template are the same logo. Pairwise samples are employed to extend the number of training samples for learning, which are collected

in an unsupervised way. Then a binary classifier is trained to further filter out false candidates and increase the detection performance.

Our contributions in this paper are summarized as follows:

i A coarse to fine framework is proposed to recognize visual logos in video streams. By this way, we reduce the false-alarm rate and improve the recognition result greatly.

ii Instead of using traditional templates, iconic templates are selected to generate stable templates of different logos to avoid the unstable phenomenon during the template initialization.

iii By introducing learning strategy with training data sampling, multiple logo matching is transformed to be a binary classification problem, which significantly improves the matching result in a fine stage.

2. RELATED WORK

The existing logo recognition approaches can be summarized into two main aspects: template matching and learning based methods. When detecting multiple logos with similar appearance in a scene, researchers often leverage on template matching methods for their simplicity and efficiency. An early approach of template matching and its extension[6] use the Chamfer distance between the template and the input image contours as a dissimilarity measure. This distance can efficiently be computed using image Distance Transform (DT). Extracted contour points are relatively fragile. Moreover, the threshold for matching is usually difficult to set when detecting multiple logos. Both HOG(Histogram Orient Gradients)[1] and SIFT(Scale-invariant Feature Transform)[2], which describe the local distributions for image gradients, are very popular used for logo recognition. They have been proven to be reliable methods but tend to be slow due to the high computational complexity. Additionally, they are difficult to handle low-textured logos. DOT(Dominant Orientation Templates)[3] is proposed for low-texture object detection and tracking, which assumes an environment with relatively simple and slow motion. The performance of these template based approaches is usually sensitive to the selection of an initial template. Xie *et al.*[7] proposed a cascade structure with gradient, texture and color information for object detection. Methods such as [8, 9, 2] usually evaluate the similarity based on the whole image and thus only require the image containing one dominate object. Wang *et al.*[10] used multispectral images gradient to find and track TV logos in video streams. Kuo *et al.*[11] used a gradient-based average approach to detect and remove logos in a video sequence. Natarajan *et al.*[12] detected multi-class logo in broadcast videos with a cascade manner, which combined color and edge based features to locate the logo template.

On the other side, learning based methods have good generalization in some aspects, e.g. logo with different orientation and with scale diversification [13, 14]. A framework for fine-grained object recognition is proposed in [15]. Laurent *et al.*[16] tried to find images containing objects that are similar to the query image via many local descriptors. These learning approaches usually need huge amounts of training samples[17, 18]. These learning based methods are seldom used in template matching due to few training samples. Huang *et al.*[19] learned a logo transition template to find logos in sports videos. T. Malisiewicz *et al.*[20] used few or even just one positive training sample to train a detector by introducing exemplar SVM, which makes detection process more efficient, it also needs complex process to mine positive samples during training. Liao *et al.*[21] first applied pairwise method by representing protein evolutionary and structural relationships by SVM-Pairwise.

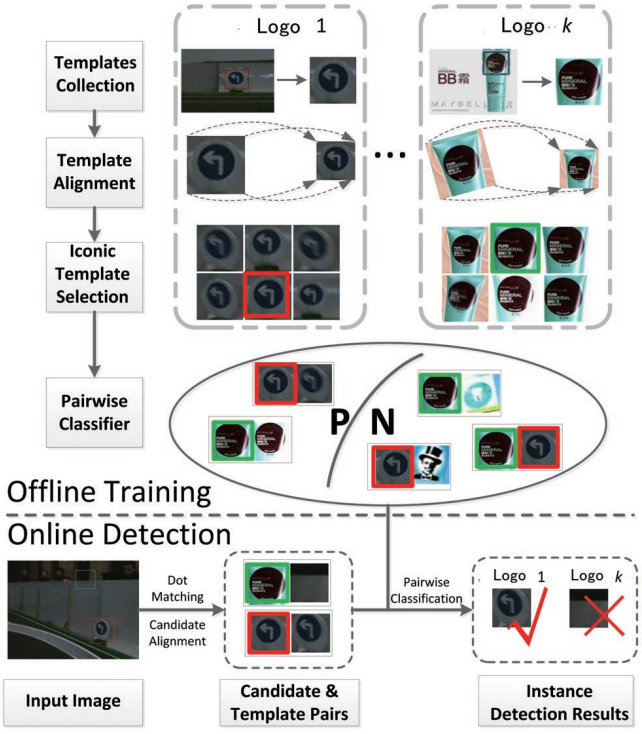

Fig. 1. Overview of the proposed approach.

3. PROPOSED APPROACH

In this paper, we propose a coarse-to-fine approach to recognize logos in video streams. Both template matching and pairwise learning approaches are combined to complement each other. The framework of our approach is shown in Fig. 1.

Our approach includes two stages: offline training and online detection.

In the training phase, an "iconic template" is selected from several initial templates. Applying the "iconic template" helps to reduce the diversity existing in different initial template selections. In addition, we use the initial templates to construct pairwise samples and learn a binary classifier. We collect many template pairs(two template images together form a pair) as training examples. Then, a pairwise classifier is trained to predict whether two images containing the same logo. In the testing phase, we firstly use the iconic template for a coarse template matching in an incoming video frame to find logo candidates. Then, the candidates and the iconic template are regarded as inputted pairs to the binary classifier. The classification result indicates whether the candidate belongs to the template logo class or not.

3.1. Template Alignment and Selection

Usually, the template is manually selected from an image containing an logo as shown in template collection part of Fig. 1. The obtained templates varied for different initialization, which may cause unstable during matching phase. To generate a robust template for each logo, here we develop a iconic template selection strategy. We firstly annotate some templates with different sizes in several images, and apply coarse template matching in the training images for templates collection. Since the obtained templates have various views and scales, we normalize them through perspective transformation. Given a template image T_k ($k \leq K$, K is the number of logos), let (x_i, y_i) denote its i^{th} pixel coordinate, and $(x_i^{'}, y_i^{'})$ the corresponding pixel in the aligned image T_{ak}. Then, the template images are aligned as follows,

$$Ta_k = ST_k \tag{1}$$

where S represents the transformation matrix, subject to the four transforms in Eq. 2. Assume the normalized template size is $M \times N$. And (x_c, y_c), (c = 0, 1, 2, 3), are the four vertices coordinates sequentially of template images. As M, N and (x_c, y_c) (c = 0, 1, 2, 3) are known, the transformation matrix S could be estimated from Eq. 2. An example of template alignments can be seen in Fig. 1.

$$\begin{pmatrix} 0 & M & M & 0 \\ 0 & 0 & N & N \end{pmatrix} = S \begin{pmatrix} m_0 & m_1 & m_2 & m_3 \\ n_0 & n_1 & n_2 & n_3 \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{2}$$

After template alignment, we select one as the iconic template for logo matching. A blurred image would be obtained if we simply average the templates, which will degrade the matching performance. Therefore, an iconic template image is selected as the representative one which could minimize the overall distance with others for the same logo *k*. For the aligned positive template images $T_{ki}, i \in I_k$ (I_k is the number of the templates for logo *k*). The iconic template T_{kr} is calculated as,

$$T_{kr} = \arg\min_i \sum_j Dis(T_{ki}, T_{kj}) \tag{3}$$

where $Dis(\cdot, \cdot)$ is the Euclidean distance in the RGB color space. These selected iconic templates are insensitive to the template initialization and improve the result of logo template matching to some extends.

3.2. Coarse template matching

All the iconic templates of different logos are used for the coarse template matching stage. Notice that any existing traditional template matching approaches can be used here, including HOG, SIFT and DOT *et al*. Though SIFT performs good on template matching, it is hard to extract enough feature points on texture less templates such as traffic signs as shown in Fig. 2, and it also suffers from computation cost. Here we adopt DOT template matching for its both accuracy and high efficiency. For a image patch that matches the iconic template, it will be treated as a logo candidate and together with the template been sent to the second stage for further refinement.

4. PAIRWISE BINARY CLASSIFICATION

For K logos, samples are usually insufficient for training K classifiers since the appearance is unchanged for one given logo. To solve the problem of insufficient positive samples, we transform the multi-classification problem to binary classification by introducing pairwise samples.

In the training phase, for each collected logo template T_i, we rotation and scaling it by perspective transformation as Eq.(2). In this way, we expand our original labeled data. As in Eq.4, each training pairwise sample consists of two logo templates or their transformations. Given K kind of logos, the i^{th} template of the k^{th} logo is denoted as T_{ki}. Analogously, the pairwise S_{klij} is constituted as $\{T_{ki}, T_{lj}\}$. And its label is given according to equation below.

$$\begin{aligned} S_{klij} &= (T_{ki}, T_{lj}) \\ L_{klij} &= \begin{cases} 1 & \text{if } k = l \\ 0 & \text{else} \end{cases} \end{aligned} \tag{4}$$

If the two templates of S_{klij} belong to the same logo, the pairwise S_{klij} is marked as positive sample for our binary classifier. Otherwise, it is a negative sample. Hence, the positive sample set is,

$$S^+ = \{S_{klij} | L_{klij} = 1.\ k, l \leq K, i, j \leq I_k\} \tag{5}$$

Similarly, the negative sample set is,

$$S^- = \{S_{klij} | L_{klij} = 0.\ k, l \leq K, i, j \leq I_k\} \tag{6}$$

where I_k is the total number of templates for the k^{th} logo.

In this way, more samples could be utilized by collecting pairs from different logos to learn one classifier, compared to train K classifiers respectively. This is an unsupervised process since the pairwise samples don't need any annotations. We then extract the pyramid local ternary pattern histogram (pLTP) [17] for each template since it is unrelated to the template size. We denote H_{ki} as the pLTP feature of template T_{ki}, and calculate the intersection histogram between H_{ki} and H_{lj} as the pair description in Eq.7.

$$D(S_{klij}) = \min\left(H_{ki}(p), H_{lj}(p)\right), p = 0, 1, \cdots, P-1 \quad (7)$$

where P is the number of histogram bins. In our case, the dimension of $D(S_{klij})$ is 1062 ($59dim \times 2parts \times 3channels \times 3scales$, details can be seen in [17]), we reduce the dimension to 64 with PCA. Then a SVM classifier is trained with $D(S_{klij})$.

In the testing phase, if we obtain a logo candidate C_{ki} by template matching with the iconic template of the k^{th} logo T_{kr}, a pairwise sample $S_{ki} = (T_{kr}, C_{ki})$ is generated by combine them together. Then the the intersection histogram features are extracted and been input into the classifier. The candidate C_{ki} is labeled as logo k or not with the classification result.

$$f(S_{ki}) > 0, \text{ where } f(S_{ki}) = \omega^T D(S_{ki}). \quad (8)$$

here $D(S_{ki})$ is the intersection histogram feature computed using Eq.7 and ω is parameter learned using SVM.

This kind of pairwise representation of two samples offers three primary advantages over the separate classifiers combination. First, the pairwise method reduces classifier number to one meanwhile increases the number of training samples. Second, after collecting some logo templates, the whole training process is in an unsupervised way. And third, this binary classifier could effectively refine the results of template matching, robust to a relatively low threshold setting. The binary classifier here is used to discriminate the similarity between the logo candidate and the iconic template, so it can be trained using pairwise samples of logos from different categories.

5. EXPERIMENT

5.1. Dataset

To evaluate the performance of the proposed approach, we collect two video dataset from two different domains, including YouTube advertising videos and traffic videos. The YouTube advertising videos contain 4 different kinds of logos including beer, camera, shampoo and toothpaste, we collect 4 different logos for each kind. The traffic videos include 4 logos like "Left", "Right", "ConeMark" and "NoParking". Only part of the traffic videos contain desired logos to be recognized. Examples of the dataset are shown in Fig. 2.

5.2. Experiment results and analysis

Recall that the coarse template matching stage can be substituted to any other template matching approaches, we choose the DOT approach[3] here for its strong ability to describe images with low texture details. We firstly evaluate the iconic template selection strategy in logo recognition. Then we compared our approach with dot matching and multiple binary classifiers approach. Multiple binary classifiers stands for training classifier for each logo.

Fig. 3 demonstrates the logo "Fosters" detection process in YouTube videos. Since our first stage detector relies on the coarse match results, we lower the DOT matching threshold in our first stage to reduce the false negative rate, the resulting more false candidates are rejected by the pairwise binaries classification stage, while the correct detection candidate been accepted shows green bounding box.

Fig. 2. Example of collected logos in YouTube videos and Traffic viedos.

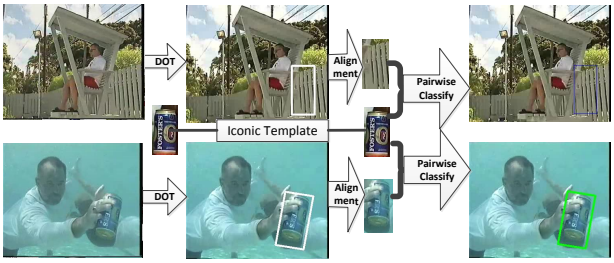

Fig. 3. Example of detect vidual logo "Foster" in YouTube videos.

We first shows the the necessity of iconic template selection strategy in the first stage of our cascade detector, here the template matching thresholds are set to the same for both iconic template and original matching approach. The ROC curve in Fig. 4 shows the result for logo "Darlie" detection in the YouTube videos. The iconic template performs better than the DOT approach, that is, the iconic template represents the logo better than a randomly chosen one and insensitive to

Table 1. Results of multiple logo detection

category	Camera				Traffic sign			
logo	Cacon	Nikon	Olympus	Fujifilm	Right	ConeMark	Left	No parking
TPR(ours)	**0.5573**	0.4481	0.4756	**0.7113**	**0.7906**	**0.9600**	**0.8062**	**0.5125**
FPR(ours)	**0.0164**	**0.0594**	**0.0122**	**0.0419**	0.1462	**0.0136**	**0.3215**	**0.1208**
TPR(DOT)	0.4584	**0.5726**	**0.6342**	0.6825	0.6104	0.8394	0.5455	0.3169
FPR(DOT)	0.0542	0.1920	0.2031	0.1220	**0.1393**	0.0928	0.4049	0.2257
category	Beer				Shampoo			
logo	Fosters	Beck	Heineken	Carlsberg	Clear	H.Shoulder	Rejoice	Palmolive
TPR(ours)	**0.7358**	**0.7219**	**0.6652**	**0.9620**	**0.7516**	**0.8933**	**0.9480**	**0.8882**
FPR(ours)	**0.0940**	**0.0903**	**0.1450**	0.0988	0.0678	0.1333	**0.0056**	**0.0085**
TPR(DOT)	0.6533	0.4956	0.2551	0.7754	0.4535	0.6000	0.9127	0.8487
FPR(DOT)	0.1008	0.1544	0.1486	**0.0052**	**0.0302**	**0.1105**	0.0566	0.0842

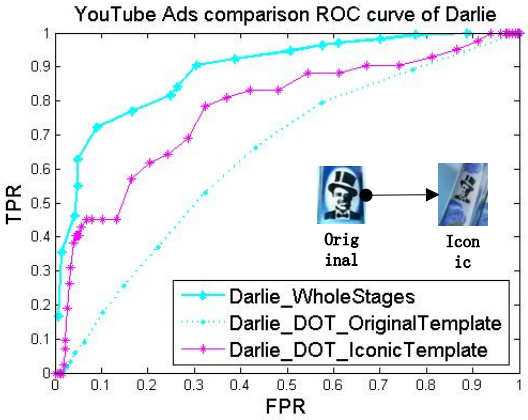

Fig. 4. ROC curves for "Iconic Template" performance.

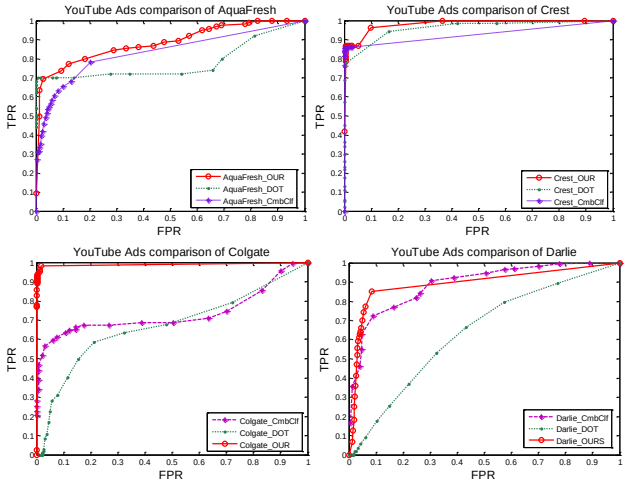

Fig. 5. Experiment ROC curves for logos in YouTube videos. "CmbClf" represents the performance of binary classifiers combination.

the template initialization.

Then we compare the over all performance of our approach with DOT matching approach[3] and multiple binary classifiers approach (to train one classifier for each logo). The ROC curves in Fig. 5 and Fig. 6 show the results in two kinds of logos both in Youtube videos and traffic sign videos, we test four logos on each kind. The results of all four logos in each kind suggest that out method outperforms the two others. We can see that multiple binary classifiers("CmbClf") shows unstable performance for different categories, we believe that is due to the inadequate of positive training samples for each classifier. we can see that some logo detection results of our approach are significantly improved than DOT, such as "Left" and "ConeMark" signs. More detail experiment result for different logos in four categories can be found in table 1, TPR and FPR mean the true positive rate and false positive rate. We can see that our method outperforms DOT significantly for "NoParking" category, that is because the logos in that category have less texture than others from different categories, DOT matching approach misses many true positive cases. our method increases the true positive rate in the first stage by lower the template matching threshold, and the generated false positive candidates are filtered out during the pairwise binary classification stage.

With frame size of 352×288, our approach takes about $30ms$ on average per frame for detecting 5 logos in the same time on a Intel i5 CPU with 4GB RAM.

6. CONCLUSION

In this paper, we propose a coarse-to-fine framework to recognize logos in video stream in real-time. Combing both coarse template matching approach and pairwise learning method together, logo recognition becomes effective and efficient through eliminating the false alarms and further refine the recognition results. Iconic template selection with template

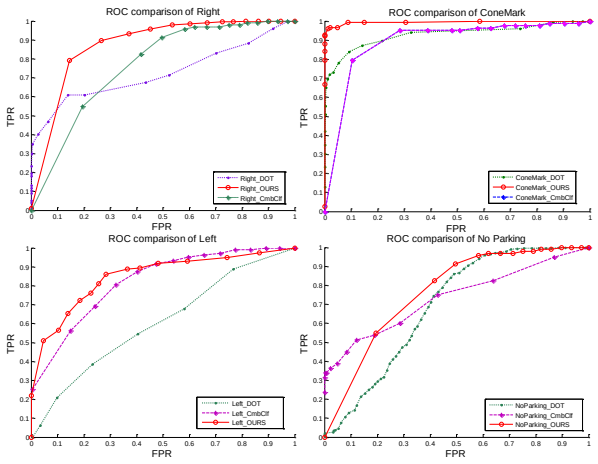

Fig. 6. Experiment ROC curves for logos in traffic videos. "CmbClf" represents the performance of binary classifiers combination.

image alignment for template matching improves the stability of the coarse stage. Experiment results show that our approach outperform the DOT matching approach and traditional multiple classifiers combination.

7. REFERENCES

[1] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005.

[2] Lowe D. G., "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[3] Hinterstoisser S., Lepetit S., Ilic S., Fua P., and Navab N., "Dominant orientation templates for real-time detection of texture-less objects," in *CVPR*. IEEE, 2010.

[4] Felzenszwalb P., McAllester D., and Ramanan D., "A discriminatively trained deformable part model," in *CVPR*. IEEE, 2008, pp. 1–8.

[5] Bay L.V., Tuytelaars H., and Gool T., "Surf: Speeded up robust features.," in *ECCV*, 2006.

[6] Gavrila D. M., "Multi-feature hierarchical template matching using distance transforms," in *ICPR*. IEEE, 1998.

[7] C. Xie, J. Wang, T. Wang, and H. Lu, "Real-time cascade template matching for object instance detection," in *PCM*, 2011.

[8] Nister D. and Stewenius H., "Scalable recognition with a vocabulary tree.," in *CVPR*. IEEE, 2006.

[9] Berg A., Berg T., and Malik J., "Shape matching and object recognition using low distortion correspondences," in *CVPR*. IEEE, 2005.

[10] Jinqiao Wang, Lingyu Duan, Zhenglong Li, Jing Liu, Hanqing Lu, and J.S. Jin, "A robust method for tv logo tracking in video streams," in *Multimedia and Expo, 2006 IEEE International Conference on*, 2006.

[11] Chung-Ming Kuo, Cheng-Ping Chao, Wei-Han Chang, and Jin-Long Shen, "Broadcast video logo detection and removing," in *Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP '08 International Conference on*, 2008.

[12] P. Natarajan, Yue Wu, S. Saleem, E. Macrostie, F. Bernardin, R. Prasad, and P. Natarajan, "Large-scale, real-time logo recognition in broadcast videos," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, 2011.

[13] Huang C., Ai H., Li Y., and Lao S., "Vector boosting for rotation invariant multi-view face detection," in *CVPR*. IEEE, 2005.

[14] Viola P. and Jones M., "Fast multi-view face detection," in *CVPR*. IEEE, 2003.

[15] Wu Y., Liu Y., Yuan Z., and Zheng N., "Iair-carped: A psychophysically annotated dataset with fine-grained and layered semantic labels for object recognition," *Pattern Recognition Letters*, vol. 33, pp. 218–226, 2012.

[16] Wu Y., Liu Y., Zheng Z., and Zheng N., "Robust object recognition in images and the related database problems," in *Multimedia Tools and Applications*.

[17] Shotton J., Johnson M., and Cipolla R., "Semantic texton forests for image categorization and segmentation," in *CVPR*. IEEE, 2008.

[18] Shotton J., Fitzgibbon A., Cook M., Sharp T., Finocchio M., Moore R., Kipman A., and Blake A., "Real-time human pose recognition in parts from a single depth image," in *CVPR*. IEEE, 2011.

[19] Qiao Huang, Jianming Hu, Wei Hu, Tao Wang, Hongliang Bai, and Yimin Zhang, "A reliable logo and replay detector for sports video," in *Multimedia and Expo, 2007 IEEE International Conference on*, 2007.

[20] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *ICCV*. IEEE, 2011.

[21] Liao L. and Nobel W.S., "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," in *Journal of Computational Biology*, 2003.