This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2713409, IEEE Transactions on Circuits and Systems for Video Technology

Latent Bi-constraint SVM for Video-based Object Recognition

Yang Liu, Minh Hoai, Mang Shao, and Tae-Kyun Kim

***Abstract*—We address the task of recognizing objects from video input. This important problem is relatively unexplored, compared with image-based object recognition. To this end, we make the following contributions. First, we introduce two comprehensive datasets for video-based object recognition. Second, we propose Latent Bi-constraint SVM (LBSVM), a maximum-margin framework for video-based object recognition. LBSVM is based on Structured-Output SVM, but extends it to handle noisy video data and ensure consistency of the output decision throughout time. We apply LBSVM to recognize office objects and museum sculptures, and we demonstrate its benefits over image-based, set-based, and other video-based object recognition.**

***Index Terms*—object recognition, video analysis, structured-output SVM.**

I. Introduction

OBJECT recognition is an important research problem in computer vision with applications in a wide range of areas, including human-computer interaction [17], intelligent surveillance [18], industrial inspection [11], robotics [8], medical imaging [13]. Because of its importance, object recognition has been extensively studied and many algorithms have been proposed. Most existing algorithms (e.g., [2], [3], [20], [23], [25], [28], [36], [40], [44]), however, are developed to recognize objects from images. They do not address the dynamics, clutter, and noisiness of video input. Only a few methods have considered videos, e.g., video-based descriptors [14], [22], [30], [34], [35], [37], [46], image-set matching [1], [21], [26], [39], [41], [42], face recognition in video [6], [7], [9], [29], and video classification [12], [19]. However, these methods are conceptually different from ours, which will be clarified in Sec. II.

The ability to recognize objects from video has many potential applications. Consider a concrete example of building a system that allows a museum's visitors to use their cell phones to recognize objects on display. In this situation, it is more beneficial and convenient to recognize objects from video input instead of images. First, many museum objects have 3D shape, and any image can only depict a single facet of an object. Thus, an image contains much less information than a video that provides multiple views of the object. Second, in a crowded museum environment, it is more convenient for a museum visitor to record a continuous video of an object, rather than to capture occlusion-free representative images of the object.

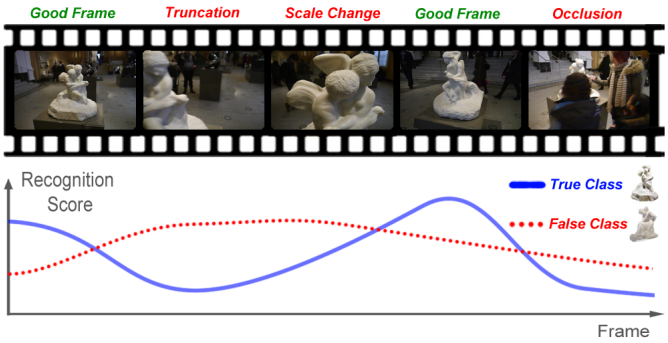


Fig. 1: **Some challenges of video-based recognition.** The noise and variation of video data cause the recognition decision of frame-based approaches to fluctuate. The object is frequently recognized as the wrong class.

Video-based object recognition, however, is challenging. Several highly important challenges are to: 1) handle the noisiness and variation of video data (e.g., not every video frame is occlusion free, and videos can vary in length, object scale, and background clutter); 2) train classifiers when relatively few video examples of each object are present; 3) effectively use the entire video for recognition and avoid the fluctuation of the recognition decision over time. Some challenges are depicted in Fig. 1.

In this paper, we propose Latent Bi-constraint SVM (LBSVM), a novel algorithm for video-based object recognition. LBSVM is built on Structured-Output SVM (SOSVM) [38], but extends it to address the challenges of recognizing objects from video input. LBSVM introduces two novel constraints and a latent variable. Its technical novelty is threefold: 1) The first constraint (Eq. 2) expands the training video, associates the object label to all subsequences of each training video. This enforces all subsequences of training video to be correctly classified, enabling the recognition of an object from various view points. It also maximizes the usage of training data, reducing the need for a large number of training videos. 2) The second constraint (Eq. 3) requires the monotonicity of the score function with respect to the inclusion relationship between subsequences of a video. This is to ensure the consistency of the recognition decisions. 3) The incorporation of the latent variable allows the monotonicity requirement to

Yang Liu, Mang Shao and Tae-Kyun Kim are with the Department of Electrical and Electronic Engineering, Imperial College London, UK. Email: {y.liu11, ms2308, tk.kim}@imperial.ac.uk

Minh Hoai is with the Department of Computer Science, Stony Brook University, SUNY, USA. Email: minhhoai@cs.stonybrook.edu

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2713409, IEEE Transactions on Circuits and Systems for Video Technology

be satisfied, discarding bad views of an object due to such factors as occlusion and motion blur. The two constraints and the latent variable allow LBSVM to ground the recognition decision on the entire video, avoiding the inconsistency of the output decisions.

We will demonstrate the benefits of LBSVM for recognizing office objects and museum sculptures from videos recorded using a handheld camera. Videos of office objects were recorded in a cluttered office environment, while museum sculptures were recorded inside a crowded museum. These *in-the-wild* videos are challenging for object recognition due to various factors, including occlusion, background clutter, scale variation, illumination change, and motion blur.

II. Related Work

A majority of algorithms for object recognition [2], [3], [20], [23], [25], [28], [36], [40] assume the input is a single image. They can be adapted to work with video input by running image-based recognition on individual frames and subsequently accumulating the recognition scores [24], [27], [31], [32]. This approach, however, has several drawbacks. First, extracting frame-level descriptors and running image-based recognition for all individual frames are inefficient; this fails to consider the temporal similarity of nearby frames. Second, a simple approach for pooling evidence from all frames can lead to poor recognition performance due to dominance of irrelevant information from frames with occlusion or motion blur. Third, this approach fails to take into account the sequential nature of video data and may produce inconsistent decisions over time.

Algorithms for object recognition in video exist. Most of them [14], [22], [34], [35], [37], [46] propose to utilize the temporal information in video and improve local video descriptors by feature tracking. [35] tracks image patches using optical flow and learns an invariant feature for recognition. [37] proposes an efficient search space for interest points to track features, which are then exploited to recognize objects. [46] proposes the RVO-SIFT method based on feature tracking for rigid video object recognition, which unifies the object recognition and feature updating process, hence improves the completeness of the video object's features automatically. [22] develops Best Template Descriptors (BTD) from video, quantizes them to generate a Bag-of-Words model, followed by a NN classifier to recognize object in video. These methods improve feature descriptors for video, but they perform object recognition frame by frame. They neither address the insufficiency of training data nor ensure the consistency of frame recognition decisions.

A video is an ordered set of images. As such, video-to-video matching can be cast as set-to-set matching [1], [21], [26], [39], [41], [42]. [26] proposes two set-of-sets representations for a video and respective matching methods for object recognition in video. The combined set-of-sets method based on bag of words and manifold techniques improves the video object recognition. [42] proposes Kernel Principal Angles (KPA), which measures the intersection of two manifolds representing two sets. Since images in a video are collected continuously, they exhibit smooth data changes and can be well constrained on a low-dimensional manifold. KPA can be used to match video manifolds. However, image-set methods are often developed especially for face recognition [1], [21], [41], including face recognition in videos [6], [29], or character identification in television shows [7], [9], which differ from the problem we tackle. A prerequisite of those works is face detection and tracking, but no detectors are available for generic objects in our case. Also, they assume temporal coherence cues, which might not hold for our videos.

In this paper, we develop LBSVM to exploit the structured information contained in video for object recognition. LBSVM is based on SOSVM [38], which can learn a correlation function between a complex input space and a structured output space. SOSVM has been shown to be widely useful in many computer vision tasks, and it has also been extended in several ways. [15] uses SOSVM for adaptive tracking and detection. [48] proposes two-layer SOSVM to recognize unsuccessful activities. [16] extends SOSVM for early event detection by anticipating the sequential nature of temporal events. SOSVM is also extended to handle latent variable [45]. [33] introduces similarity constraints for weakly supervised action classification, which performs classification and discriminative localization. [43] utilizes kernelized SOSVM for recognizing human actions from arbitrary views, which implicitly infers the view label by latent variable during both training and testing. Another way to handle latent variable is by Multiple Instance Learning [47], but it can neither exploit the structured information in video to ensure consistent decision, nor be successfully trained when relatively few video examples per class are present.

LBSVM learns and recognizes class labels of videos. It is different from [14], [22], [34], [35], [37], which perform frame-by-frame recognition in video. It is also different from previous works requiring finer-level annotation, such as [4], which propagates the pixel label of initial frame through video.

III. Latent Bi-Constraint SVM

A. Learning Formulation

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$ be a set of training videos. Each video $\mathbf{x}_i$ depicts an object, and let y_i be the label of that object. Let $\{\mathbf{x}_i^t\}$ be the set of all subsequences of video $\mathbf{x}_i$, at all locations and scales, as illustrated in Fig. 2. We learn a LBSVM for video-based object recognition by solving the following optimization problem:

$$\underset{\mathbf{w}}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||^2 \tag{1}$$

$$\text{s.t. } f_{\mathbf{w}}(\mathbf{x}_i^t, y_i) - f_{\mathbf{w}}(\mathbf{x}_i^t, y) \geq 1 \ \forall i, \forall t, \forall y \neq y_i, \tag{2}$$

$$f_{\mathbf{w}}(\mathbf{x}_i^t, y_i) - f_{\mathbf{w}}(\mathbf{x}_i^j, y_i) \geq \Delta(\mathbf{x}_i^t, \mathbf{x}_i^j) \quad \forall i, \forall t, \forall j : \mathbf{x}_i^j \subset \mathbf{x}_i^t. \tag{3}$$

Here $f_{\mathbf{w}}(\mathbf{x}, y)$ is the score function for a video segment $\mathbf{x}$ and a label y. We consider a linear recognition score function $f_{\mathbf{w}}(\mathbf{x}, y) = \mathbf{w}^T \psi(\mathbf{x}, y)$. $\psi(\mathbf{x}, y)$ is the joint feature mapping of the video segment $\mathbf{x}$ and the label y, and $\mathbf{w}$ is the parameter of the score function, which needs to be learned. Constraint (2)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2713409, IEEE Transactions on Circuits and Systems for Video Technology

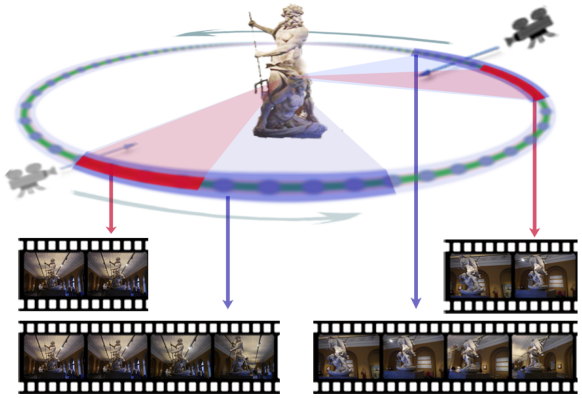


Fig. 2: Subsequences of a training video at different locations and scales. See Sec. IV-B2 *Subsequences* for details.

Fig. 3: Using a latent variable, LBSVM removes bad views to avoid the fluctuation of the recognition score, keeping it increasing.

requires all subsequences of a training video to be correctly classified. This is based on the observation that if a video depicts an object then all subsequences of the video also depict the same object, possibly from different angles. This constraint essentially trains the system to recognize various views of the object. Constraint (3) requires monotonicity of the recognition function with respect to the inclusion relationship between video subsequences—the recognition score (confidence) of a video segment should not be lower than the recognition score of its subsequences. This constraint reflects a fact that a long video segment has more views of the object than a shorter one, and therefore, it should be recognized with higher confidence, as illustrated in Fig. 3. In Constraint (3), $\Delta(\cdot,\cdot)$ is the adaptive margin, and we use $\Delta(\mathbf{x}_i^t, \mathbf{x}_i^j) = 1 - \frac{|\mathbf{x}_i^j|}{|\mathbf{x}_i^t|}$. By optimizing Eq. (1), we obtain the function $f_{\mathbf{w}}$, that can be used for video-based object recognition. Given a new testing video $\mathbf{x}$, we predict the label of the object it depicts by finding the label that maximizes the score:

$$y^* = \underset{y}{\operatorname{argmax}}\, f_{\mathbf{w}}(\mathbf{x}, y). \tag{4}$$

Here, $\mathbf{x}$ is the entire video, which contains more views of the object than any of its subsequences. As in the traditional formulation of SVM, the constraints are allowed to be violated by introducing slack variables:

$$\begin{aligned} \underset{\mathbf{w},\alpha_i^t,\beta_i^t}{\text{minimize}}\ & \frac{1}{2}||\mathbf{w}||^2 + C_1 \sum_{i=1}^{m}\sum_{t} \alpha_i^t + C_2 \sum_{i=1}^{m}\sum_{t} \beta_i^t \qquad (5)\\ \text{s.t. } & f_{\mathbf{w}}(\mathbf{x}_i^t, y_i) - f_{\mathbf{w}}(\mathbf{x}_i^t, y) \geq 1 - \alpha_i^t \ \forall i, \forall t, \forall y \neq y_i,\\ & f_{\mathbf{w}}(\mathbf{x}_i^t, y_i) - f_{\mathbf{w}}(\mathbf{x}_i^j, y_i) \geq \Delta(\mathbf{x}_i^t, \mathbf{x}_i^j) - \beta_i^t\\ & \qquad\qquad \forall i, \forall t, \forall j : \mathbf{x}_i^j \subset \mathbf{x}_i^t,\\ & \alpha_i^t \geq 0, \beta_i^t \geq 0 \ \forall i, \forall t. \end{aligned}$$

Although a video primarily depicts an object, the video might also contain not-so-informative frames due to several factors such as background clutter, occlusion, and motion blur. To filter out these irrelevant frames, we introduce latent variables into the feature mapping and the recognition function:

$$f_{\mathbf{w}}(\mathbf{x}, y) = \max_{h} \mathbf{w}^T \psi(\mathbf{x}, y, h). \tag{6}$$

For a video sequence $\mathbf{x}$ that is represented by l frames, h is an indicator vector for selecting representative views of the object, and $\psi(\mathbf{x}, y, h)$ computes the feature representation on selected frames. As illustrated in Fig. 3, where the use of the indicator h is necessary to filter out bad views of the object to avoid the fluctuation of the recognition score. Details about the selector vector h and the feature representation will be described in Sec. IV-B.

Once the model parameter $\mathbf{w}$ has been learned, inferring the label for a test video is done jointly with selecting good views of the object in the video:

$$(y^*, h^*) = \underset{y,h}{\operatorname{argmax}}\, \mathbf{w}^T \psi(\mathbf{x}, y, h). \tag{7}$$

By exploiting the temporal similarity of nearby frames in a video, we can limit the domain of h and only consider frames at a regular interval. This allows the above joint inference problem to be solved efficiently using exhaustive search. This will be described in more details in Sec. IV-B.

B. Optimization

The constrained optimization in Eq. (5) is equivalent to the following unconstrained problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + R(\mathbf{w}), \tag{8}$$

where

$$R(\mathbf{w}) = \sum_{i=1}^{m}\sum_{t} \left(C_1 \max\{0, R_{it}^1\} + C_2 \max\{0, R_{it}^2\}\right),$$

and R_1^{it} and R_2^{it} are:

$$\begin{aligned} R_{it}^1 &= \max_{y \neq y_i} f_{\mathbf{w}}(\mathbf{x}_i^t, y) + 1 - f_{\mathbf{w}}(\mathbf{x}_i^t, y_i),\\ R_{it}^2 &= \max_{j:\mathbf{x}_i^j \subset \mathbf{x}_i^t} [f_{\mathbf{w}}(\mathbf{x}_i^j, y_i) + \Delta(\mathbf{x}_i^t, \mathbf{x}_i^j)] - f_{\mathbf{w}}(\mathbf{x}_i^t, y_i) \end{aligned}$$

We use Non-convex Regularized Bundle Method (NRBM) [10] to optimize for Eq. (8). NRBM iteratively constructs an increasingly accurate piecewise quadratic lower bound of the objective function. In each iteration, a new cutting plane is found by the sub-gradient of the objective

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2713409, IEEE Transactions on Circuits and Systems for Video Technology

function and added to the piecewise quadratic lower bound approximation. The algorithm starts from a random $\mathbf{w}_1$ and generates a sequence of $\mathbf{w}_i$'s. The algorithm terminates when the gap between the minimum of the approximation function and the value of the objective function is smaller than a predefined tolerant value.

The sub-gradient of $R(\mathbf{w})$ w.r.t. $\mathbf{w}$ can be computed from the gradients of R^1 and R^2. From the linear form in Eq. (6), the subgradient $\frac{\partial R(\mathbf{w})}{\partial \mathbf{w}}$ is:

$$\sum_{i=1}^{m}\sum_{t}\Big\{C_1[\psi(\mathbf{x}_i^t, y_{it}^1, h_{it}^1) - \psi(\mathbf{x}_i^t, y_i, h_{it})]H(R_{it}^1) + C_2[\psi(\mathbf{x}_i^{t'}, y_i, h_{it}^2) - \psi(\mathbf{x}_i^t, y_i, h_{it})]H(R_{it}^2)\Big\} \quad (9)$$

where $H(\cdot)$ is the Heaviside step function:

$$H(R_{it}) = \begin{cases} 1 & \text{if } R_{it} \geq 0 \\ 0 & \text{if } R_{it} < 0, \end{cases}$$

and (y_{it}^1, h_{it}^1), $(x_i^{t'}, h_{it}^2)$, h_{it} are inferred by:

$$(y_{it}^1, h_{it}^1) = \underset{y \neq y_i, h}{\operatorname{argmax}} \mathbf{w}^T \psi(\mathbf{x}_i^t, y, h), \quad (10)$$

$$(\mathbf{x}_i^{t'}, h_{it}^2) = \underset{j:\mathbf{x}_i^j \subset \mathbf{x}_i^t, h}{\operatorname{argmax}} [\mathbf{w}^T \psi(\mathbf{x}_i^j, y_i, h) + \Delta(\mathbf{x}_i^t, \mathbf{x}_i^j)], \quad (11)$$

$$h_{it} = \underset{h}{\operatorname{argmax}} \mathbf{w}^T \psi(\mathbf{x}_i^t, y_i, h). \quad (12)$$

In each iteration of NRBM, we infer (y_{it}^1, h_{it}^1), $(x_i^{t'}, h_{it}^2)$, h_{it} and optimize model parameter $\mathbf{w}$ respectively:

1) Fix the model parameter $\mathbf{w}$, infer (y_{it}^1, h_{it}^1), $(x_i^{t'}, h_{it}^2)$, h_{it} by Eqs. (10–12).
2) Fix (y_{it}^1, h_{it}^1), $(x_i^{t'}, h_{it}^2)$, h_{it}, finding a new cutting plane by Eq. (9) and add it to the quadratic piecewise approximation of NRBM, updating the model parameter $\mathbf{w}$ by minimizing the quadratic approximation.

The learning process is shown in Algorithm 1.

Algorithm 1: LBSVM Learning

Input: $\{\mathbf{x}_i, y_i\}_{i=1}^m$: training videos and labels, C_1 and C_2: slack variable coefficients, ϵ: gap threshold
Output: Model parameter $\mathbf{w}$

```
Initialize the model parameters w_1 randomly
for i ← 1 to m do
    Generate all subsequences x_i^t of x_i
end
while true do
    for i ← 1 to m, each t do
        - Fix w, infer latent variables (y_it^1, h_it^1), (x_i^t', h_it^2),
          h_it by Eqs. (10–12)
        - Given updated (y_it^1, h_it^1), (x_i^t', h_it^2), h_it,
          recompute the feature representation for each
          video subsequence
    end
    Compute subgradient c_w = ∂R(w)/∂w from Eq. (9)
    Update w by Eq. (13) of NRBM [10]
    Compute w* and gap by Algorithm 1 of NRBM
    if gap < ε then
        break;
    end
end
return w*
```

IV. Experiments

This section introduces two datasets for video-based object recognition and demonstrates the benefits of LBSVM over frame-based, set-based, and video-based approaches.

A. Datasets

*1) **Office Dataset**:* The *Office* dataset contains 210 videos of 10 object categories in a cluttered office environment: mouse, keyboard, fan, monitor, computer case, chair, pen holder, headset, stapler, scissor. Some example frames are shown in Fig. 4(a). Each object category contains videos of 5 object instances (see Fig. 4(c)). Training data is a video spanning 360° of one instance (Fig. 2). Testing data is the remaining 20 videos of the other 4 instances, recorded with different variations (e.g., Fig. 4(d)). In total, there are 10 training videos and 200 testing videos. The durations of training videos are approximately 15 seconds, and the lengths of testing videos range from 6 to 10 seconds. 100 frames were extracted from each training video, and 20 frames per second were extracted from each testing video. The spatial resolution of all videos is 640×480 pixels. The *Office* dataset is challenging, with heavy clutters, extreme scales, illumination changes, and view shifting. In some frames, the object is even out of sight. Some challenging images are shown in Fig. 4(b). The variation of an object in a video is shown in Fig. 4(d).

*2) **Museum Dataset**:* The *Museum* dataset contains 820 videos of 20 sculptures. The sculptures are 3D objects with low texture, and many of them have similar appearance. The sculptures includes portrait miniatures, statues, busts, as shown in Fig. 5(a). Each sculpture has 41 videos: one is used for training and 40 for testing. The testing data is further divided into two equal and disjoint subsets, called Museum1 (easy) and Museum2 (hard). In total, there are 20 videos for training, 400 testing videos in Museum1 and the other 400 testing videos in Museum2. The videos in the training set, Museum1, Museum2 last for around 20 seconds, 6-10 seconds, and 5 seconds respectively. The videos of the Museum dataset have the same format as the *Office* dataset, including frame rate and spatial resolution. All videos were captured by a handheld camera.

All videos were collected during rush hours, when the museum was crowded and the sculptures were surrounded by many people, the occlusion of the target sculpture and the background clutter were heavy. The training videos were taken by moving the camera around each sculpture. The testing videos were collected by imitating the habits of average users. There are significant scale changes in the dataset. Some videos only partially cover the sculptures in the close distance, some other videos capture the sculptures as one tenth of the view. Most users are inexperienced photographer, the target sculpture is often not in the middle of the view, it sometimes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2713409, IEEE Transactions on Circuits and Systems for Video Technology

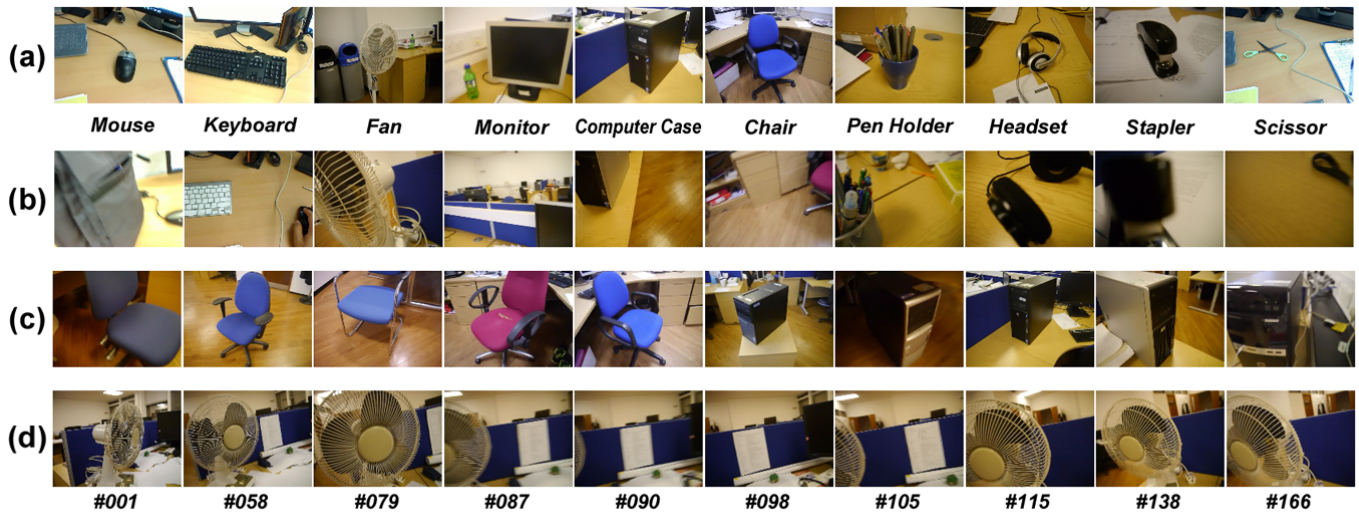


Fig. 4: Example frames of *Office* dataset: (a) representative frames of the 10 object categories; (b) challenging frames of the 10 categories; (c) the 5 instances in the category of chair and computer case; (d) the variations of a fan in a video.

slips out of the camera's point of view. Some challenging frames are shown in Fig. 5(b).

B. Feature Representation

1) ***Feature extraction****:* For feature extraction, we use Dense SIFT (DSIFT) [3], [28]. Subsequently, Spatial Pyramid Matching with Sparse Coding (ScSPM) [44] and Bag-of-Words (BoW) [34] were used for aggregating descriptors for the *Office* dataset and the *Museum* dataset respectively. These representations are common for all methods in the experiments.

For ScSPM in *Office* dataset, all settings followed the standard way in [44]. DSIFT was extracted from patches of 16×16 pixels from each sampled image with step size of 6 pixels. The codebook size was 1024, and we used spatial pyramid with 3 levels. The descriptors were aggregated using maximum pooling. The feature dimension was reduced to 150 by PCA.

For BoW in the *Museum* dataset, DSIFT was extracted at every 4 pixels with 4 patch sizes, 16×16, 24×24, 32×32 and 40×40 pixels. The codebook size was set as 300. The coded descriptors were aggregated by average pooling.

2) ***Subsequences****:* We generated subsequences for each training video as follows. First, for each training video, we extracted 100 frames. The subsequences are sampled at 10 scales (from 10 to 100 frames) and at a regular interval (every second frame). In total, we generated 235 subsequences for each training video. These sequences correspond to multiple views of an object (Fig. 2).

3) ***Subsequence representation****:* From each subsequence, l frames were uniformly sampled. A subsequence is represented as a ScSPM or BoW feature vector, by pooling all quantized descriptors from the sampled frames. This could include the noise descriptors from the bad frames. To deal with this problem, we introduced a latent variable h, which selected half (empirically set, fixed in all experiments) of the l frames, and encoded all possible selections as values of h. For each value of h, we computed a joint feature $\psi(\mathbf{x}, y, h)$ (ScSPM or BoW) by pooling the quantized descriptors from the selected frames, rather than from all the frames.

The temporal similarity of nearby frames in a video allows the subsequences to be subsampled. This limits the domain of the latent variable h to be less than hundreds. Since a linear model is used in our algorithm, it is feasible to cope with all possible values of the latent variable.

C. Compared Methods

We compared the proposed method with several frame-based, set-based, and video-based recognition methods. Detailed parameter settings of these methods are given in Secs. IV-D and IV-E.

1) ***Frame-based recognition (Frame)****:* This method takes a frame, i.e., ScSPM [44] or BoW [34] feature vector, as the input of the classifier in both training and testing. A multi-class linear SVM is trained from the extracted frames of training videos. Each frame of testing videos is independently evaluated by the classifier. Here, we use LIBSVM [5] with one-vs-one setting.

2) ***Accumulating frame recognition (Accum-Frame)****:* This method accumulates the frame-based recognition results (same as above) of all frames in a video to vote the video class [32]. Three voting schemes are considered: Hard, Soft, and KNN voting. Hard voting uses the label results of the SVM. Soft voting adopts the probability estimation of the SVM. KNN voting selects K (empirically set as 20) frames with the best SVM scores to vote for the video class.

3) ***Best Template Descriptor (BTD)****:* BTD [22] learns video-based descriptors by feature tracking in training videos and uses a BoW model and a nearest neighbor classifier to recognize object in video frame by frame. Following [22], we learn BTD descriptors from all subsequences of training

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2713409, IEEE Transactions on Circuits and Systems for Video Technology

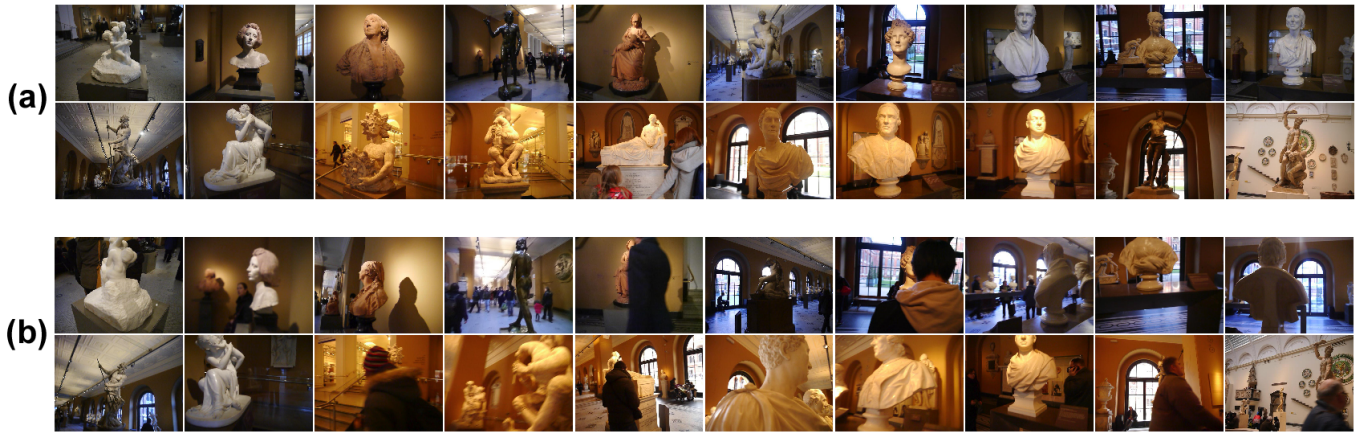


Fig. 5: Example frames of *Museum* dataset: (a) representative frames of the 20 sculptures; (b) challenging frames of the 20 sculptures.

videos, generating ScSPM (*Office*) or BoW (*Museum*) representations for them. A multi-class linear SVM classifier is trained from the subsequences, and frame-based recognition is performed for testing videos. Finally, soft voting is used to report the video-based recognition result.

4) ***Set-to-set matching (KPA)****:* Video-based object recognition can be solved by image-sets matching. We use Kernel Principal Angles (KPA) [42] to perform set-to-set matching between two videos (image sets). KPA takes two image sets as input, learns a manifold for each set, and compute the principal angles as the similarity between the two sets. Finally, a nearest neighbor classifier is used to classify a test video based on the similarity measurement.

D. Results on the Office Dataset

The second column of Tab. I shows the results of the various methods on the *Office* dataset. Frame with ScSPM, which is one of the state-of-the-art representations for image categorization, only achieves 65.1% accuracy with 100 training images per category. For a comparison, ScSPM achieves 73.2 % accuracy on the Caltech-101 dataset with only 30 training images per category [44]. This demonstrates the challenges of the *Office* dataset. The result of Accum-Frame in Tab. I is by soft voting; Accum-Frame with hard voting and KNN voting achieve lower accuracies of 73.5% and 72%, respectively. All of these results are significantly better than the result of Frame. This indicates the importance of accumulating information in a video for object recognition. BTD is the video-based descriptor method using feature tracking. It slightly improves Accum-Frame. This is perhaps because the dataset was collected by a handheld camera producing short-range egocentric view, in which objects continuously shift and move out of sight, making feature tracking unstable. KPA considers a video as a manifold and video recognition as manifold-to-manifold matching. This method yields better result than BTD and Accum-Frame. While BTD and Accum-Frame only rely on the available data, manifold estimates the unseen data by interpolation and therefore has better generalization property. Both frame-based and aforementioned video-based recognition approaches, however, are inferior to LBSVM. The better accuracy of LBSVM can be credited to its ability to make use of all subsequence information from a video and at the same time it filters out bad views of the object in the video by latent variable. Fig. 6 displays the example of selected views by LBSVM.

TABLE I: **Results on the *Office* and *Museum* datasets.** The same feature representation is used on each dataset: ScSPM [44] for the *Office* dataset and BoW [34] for the *Museum* dataset. The proposed method LBSVM achieves the best accuracy on all datasets.

Algorithm	Office	Museum1	Museum2
Frame	65.1	67.3	56.7
Accum-Frame	74.5	91.5	73.5
BTD [22]	76.0	91.8	75.3
KPA [42]	79.5	95.0	85.8
LBSVM (proposed)	**84.5**	**98.8**	**91.5**

Fig. 6: The selected good views of fan by LBSVM.

The confusion matrices of BTD and LBSVM for recognizing objects from the *Office* dataset are shown in Fig. 7. LBSVM outperforms BTD on most objects, yielding an accuracy of more than 80% for all objects, except for *mouse* and *stapler*. This might be due to the low texture appearance of mouse and stapler objects.

To analyze the consistency of the recognition decision, we evaluate the recognition accuracy over time, by running the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2713409, IEEE Transactions on Circuits and Systems for Video Technology

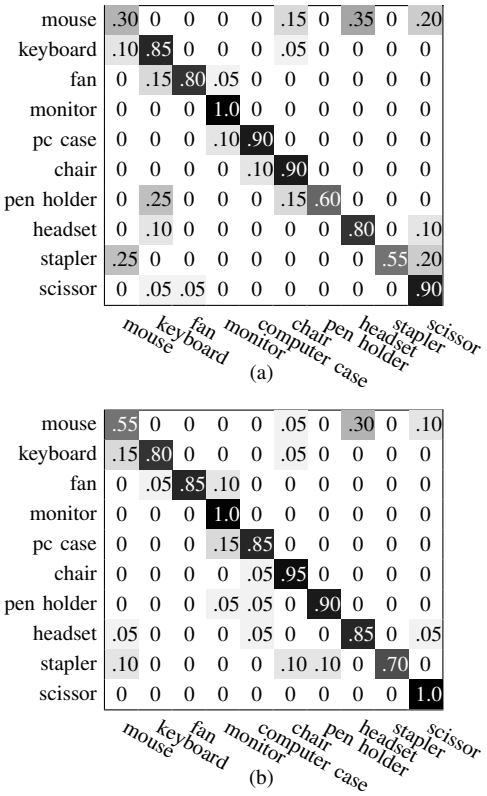


Fig. 7: Confusion matrices for recognition on the *Office* dataset: (a) BTD; (b) LBSVM.

TABLE II: **Comparison with variant methods.** BSVM is LBSVM without latent variables. SCSVM is BSVM without enforcing monotonicity constraint (Constraint (3)).

Algorithm	Office	Museum1	Museum2
SCSVM	80.0	94.5	86.0
BSVM	82.0	96.8	89.3
LBSVM (proposed)	**84.5**	**98.8**	**91.5**

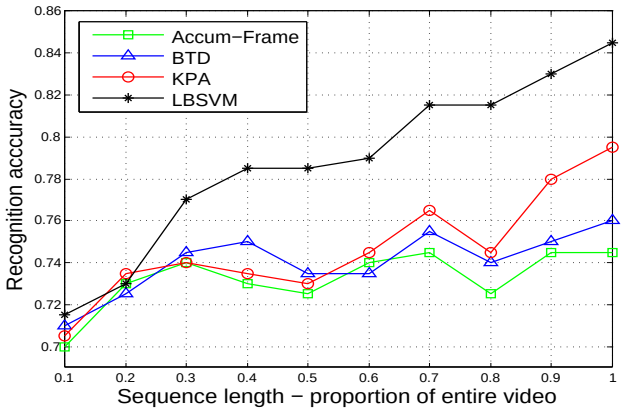


Fig. 8: The proposed LBSVM can accumulate information effectively and keep the recognition accuracy increasing monotonically, while the results of Accum-Frame, BTD, and KPA fluctuate.

recognition algorithm on subsequences of testing videos. Fig. 8 plots the recognition accuracy against the length of video subsequence (from 10% to 100% of testing videos). As the sequences become longer and more views appear, the results of Accum-Frame, BTD and KPA methods fluctuate, while the proposed method can accumulate information effectively and keep the recognition accuracy increasing. Especially in some intervals, where the compared methods decrease dramatically, the recognition performance of the proposed method still increases or remain the same. Hence, LBSVM can ground the recognition on the entire video.

Tab. II reports the performance of two variants of LBSVM. BSVM is LBSVM without the ability to discard uninformative frames. As can be seen, it does not perform as well as LBSVM. This emphasizes the importance of latent variable and view selection. SCSVM is BSVM without the monotonicity constraint (Eq. 3), and it has even lower recognition accuracy.

LBSVM is efficient. In training, using PCA for reducing the dimension of feature vectors, it took about 2 hours for *Office* dataset with the maximum iteration of 300. In testing, it took 11ms to classify a video. This excludes the time for feature extraction, which is common for all methods. Since LBSVM needs to compute features in fewer sampled frames, the time for feature extraction is also largely reduced. These timing figures were measured on an Intel Core i7 3.4GHZ×8 processor with 8GB RAM, for a Matlab implementation of LBSVM.

The parameters of the compared methods were set to report the best accuracies. Specifically, in KPA method, the Gaussian kernel was used with the bandwidth parameter $\gamma = 1$, and using the first principal angle as similarity got the best result. In SCSVM, the slack variable coefficient was set as $C = 10^{-4}$. In BSVM, the slack variable coefficients were $C_1 = C_2 = 0.5 \times 10^{-4}$. For LBSVM, the slack variable coefficients were set the same as those in BSVM. The number of sampled frames for view selection was $l = 10$. The stopping criterion for SCSVM, BSVM and LBSVM was $\epsilon = 0.01$.

E. Results on the Museum Dataset

The results of LBSVM and several other methods on the *Museum* dataset are shown in the last two columns of Tab. I. The low results of Frame show the challenges of the two *Museum* datasets. Accum-Frame (using soft-voting) significantly outperforms Frame. The other voting schemes, hard-voting and KNN-voting, do not perform as well, achieving 90.0% and 86.5% on Museum1, and 71.5% and 71.5% on Museum 2. BTD, based on video descriptors, is slightly better than Accum-Frame. KPA (manifold matching) and SCSVM (learning on all subsequences) achieved comparable performance, but were outperformed by BSVM (ensuring the consistency of recognition decision). LBSVM, by view selection and information accumulation, yielded the best results on both datasets.

The Gaussian kernel of KPA method had the bandwidth parameter $\gamma = 0.9$. The slack variable coefficients for SCSVM

and LBSVM (and also BSVM) were set as $C = 10^{-5}$ and $C_1 = C_2 = 0.5 \times 10^{-5}$. Other parameters were the same as *Office* dataset in Section IV-D.

V. Conclusions

We proposed two new datasets and a novel algorithm, LBSVM, for video-based object recognition. LBSVM is based on Structured-Output SVM, but extends it to handle noisy video data and ensure consistency of the output decisions. LBSVM introduces two novel constraints. The first constraint expands training videos and requires all subsequences to be correctly classified, training the classifier to recognize testing videos of various views. The second constraint imposes monotonicity of the score function with respect to the inclusion relationship between subsequences of a video. Furthermore, LBSVM incorporates latent variables for view selection, filtering out bad views of an object in a video. The latent variable, together with the two novel constraints, allow LBSVM to ground the recognition decision on the entire video, avoiding the inconsistency of the output decisions. In training, we optimized the parameters of an LBSVM and the latent variables iteratively. In testing, we jointly inferred the latent variable and the class label to maximize the score function. We showed that our algorithm outperformed frame-based, set-based, and other video-based object recognition approaches on the two new datasets for video-based object recognition.

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, 2005.

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[3] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[4] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2009.

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011.

[6] N. Cherniavsky, I. Laptev, J. Sivic, and A. Zisserman. Semi-supervised learning of facial attributes in video. In *First International Workshop on Parts and Attributes, in conjunction with ECCV*, 2010.

[7] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *ICCV*, 2011.

[8] A. Collet, D. Berenson, S. S. Srinivasa, and D. Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *ICRA*, 2009.

[9] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *CVPR*, 2009.

[10] T.-M.-T. Do and T. Artières. Large margin training for hidden Markov models with partially observed states. In *ICML*, 2009.

[11] E. Dominguez, C. Spinola, R. M. Luque, E. J. Palomo, and J. Muoz. Object recognition and inspection in difficult industrial environments. In *ICIT*, 2006.

[12] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[13] G. T. Flitton, T. P. Breckon, and N. M. Bouallagu. Object recognition using 3d sift in complex ct volumes. In *BMVC*, 2010.

[14] V. Gouet-Brunet and B. Lameyre. Object recognition and segmentation in videos by connecting heterogeneous visual features. *Computer Vision and Image Understanding*, 2008.

[15] S. Hare, A. Saffari, and P. H. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011.

[16] M. Hoai and F. De la Torre. Max-margin early event detectors. In *CVPR*, 2012.

[17] A. Jaimes and N. Sebe. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, 2007.

[18] O. Javed and M. Shah. Tracking and object classification for automated surveillance. In *ECCV*, 2002.

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

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

[21] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, 2003.

[22] T. Lee and S. Soatto. Video-based descriptors for object recognition. *Image and Vision Computing*, 2011.

[23] D. Levi and A. B. Hillel. Vision-based object detection by part-based feature synthesis, May 13 2014. US Patent 8,724,890.

[24] B. Li, R. Chellappa, Q. Zheng, and S. Z. Ser. Model-based temporal object verification using video. *IEEE Tran. on Image Processing*, 2001.

[25] F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[26] Y. Liu, Y. Jang, W. Woo, and T.-K. Kim. Video-based object recognition using novel set-of-sets representations. In *CVPRW*, 2014.

[27] Y. Liu, R. Kouskouridas, and T.-K. Kim. Video-based object recognition with weakly supervised object localization. In *ACPR*, 2015.

[28] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[29] S. Nagendra, R. Baskaran, and S. Abirami. Video-based face recognition and face-tracking using sparse representation based categorization. *Procedia Computer Science*, 2015.

[30] N. Noceti, E. Delponte, and F. Odone. Spatio-temporal constraints for on-line 3d object recognition in videos. *Computer Vision and Image Understanding*, 2009.

[31] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010.

[32] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *CVPR Workshops*, 2009.

[33] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori. Similarity constrained latent support vector machine: an application to weakly supervised action classification. In *ECCV*, 2012.

[34] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[35] D. Stavens and S. Thrun. Unsupervised learning of invariant features using video. *In CVPR*, 2010.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[37] D.-N. Ta, W.-C. Chen, N. Gelfand, and K. Pulli. Surftrac: Efficient tracking and continuous object recognition using local feature descriptors. In *CVPR*, 2009.

[38] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, and Y. Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 2005.

[39] S. Wan and J. Aggarwal. Robust object recognition in rgb-d egocentric videos based on sparse affine hull kernel. In *CVPR Workshops*, 2015.

[40] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[41] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, 2008.

[42] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 2003.

[43] X. Wu and Y. Jia. View-invariant action recognition using latent kernelized structural svm. In *ECCV*, 2012.

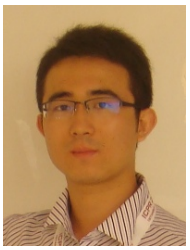
[44] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[45] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, 2009.

[46] J. Yu, F. Zhang, and J. Xiong. An innovative sift-based method for rigid video object recognition. *Mathematical Problems in Engineering*, 2014.

[47] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *NIPS*, 2005.

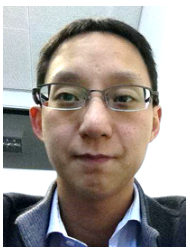
[48] Q. Zhou and G. Wang. Learning to recognize unsuccessful activities using a two-layer latent structural model. In *ECCV*, 2012.



Yang Liu received his Ph.D. degree from the Electrical and Electronic Engineering Department of Imperial College London in 2016. Prior to his PhD, he received his MSc degree from the Institute of Automation, Chinese Academy of Sciences in 2011, and BSc degree from the School of Automation Science and Electrical Engineering, Beihang University in 2008. His areas of interest include computer vision and machine learning.

Minh Hoai Nguyen is an Assistant Professor of Computer Science at Stony Brook University. He received a Bachelor of Software Engineering from the University of New South Wales in 2005 and a Ph.D. in Robotics from Carnegie Mellon University in 2012. His research interests are in computer vision and machine learning, especially human action and activity recognition and prediction.

Mang Shao is a PhD candidate in the Imperial Computer Vision and Learning Lab at Imperial College London. Prior to his PhD, he received the BSc and MEng degrees from the Electrical and Electronic Engineering Department of Imperial College London in 2012. His research interests include object recognition and 3D object pose estimation.



Tae-Kyun (T-K) Kim is an Assistant Professor and leader of Computer Vision and Learning Lab at Imperial College London, UK, since Nov 2010. He received the B.Sc. and M.Sc. degrees from Korea Advanced Institute of Science and Technology in 1998 and 2000, respectively, and worked at Samsung Advanced Institute of Technology in 2000-2004. He obtained his PhD from Univ. of Cambridge in 2008 and Junior Research Fellowship (governing body) of Sidney Sussex College, Univ. of Cambridge for 2007-2010. His research interests primarily lie in decision forests (tree-structure classifiers) and linear methods for: articulated hand pose estimation, face analysis and recognition by image sets and videos, 6D object pose estimation, active robot vision, activity recognition and object detection/tracking. He has co-authored over 40 academic papers in top-tier conferences and journals in the field, his co-authored algorithm for face image retrieval is an international standard of MPEG-7 ISO/IEC.