Video-based Object Recognition with Weakly Supervised Object Localization

Yang Liu Rigas Kouskouridas Tae-Kyun Kim
Department of Electrical and Electronic Engineering, Imperial College London
Exhibition Road, South Kensington, London, SW7 2AZ
{y.liu11, r.kouskouridas, tk.kim}@imperial.ac.uk

Abstract

With the number of videos growing rapidly in modern society, automatically recognizing objects from video input becomes increasingly pressing. Videos contain abundant yet noisy information, with easily obtained video-level labels. This paper targets the problem of video-based object recognition, whilst keeping the advantages of videos. We propose a novel algorithm, which only utilizes the weak video-level label in training, iteratively updating the classifier and inferring the object location in each video frame. During testing we obtain more accurate recognition results by inferring the location of the object in the scene. The background and temporal information are also incorporated in the model to improve the discriminability and consistency of recognition in video. We introduce a novel and challenging YouTube dataset to demonstrate the benefits of our method over other baseline methods.

1. Introduction

With the fast development of video collection and storage techniques, our lives have become filled with all kinds of videos. The number of videos uploaded on the Internet such as YouTube is increasing exponentially. The numerous monitoring cameras all over the world working for 24 hours everyday also produce countless surveillance footage. The demand of automatically understanding the semantic content of a video, such as recognizing the object in it is stronger than ever. However, compared with the large amount of research on image-based object recognition [4, 17, 18, 26] , video-based object recognition is relatively underexplored.

Video-based object recognition has its particular advantages. Firstly, a video typically contains multiple views of the target object. It possesses much more information than a single image and is prone to yielding better recognition result. Secondly, working with videos saves a large amount of annotation effort. A training video consisting of hundreds of images only requires one label, which is much less effort than labelling these images, let alone the finer bounding box annotation for image-based detection problem.

Video-based object recognition, however, also raises new challenges. Videos, especially those on the Internet, are totally unconstrained. They vary in style, length, quality, resolution, etc. They are often taken from the real-world wild environments, which bring in all kinds of illumination conditions and cluttering background. The scene shifts and changes frequently resulting in: sometimes no target object in it, sometimes multiple target objects in it. The target object in a video also varies a lot: sometimes it is as small as a dot; sometimes it is so big and only partially shown in the scene. Due to the aforementioned issues, the information in a video is very noisy, making video-based object recognition very challenging.

In this paper, we aim at solving the problem of video-based object recognition, whilst keeping the advantages of more information and less annotation effort of videos. Note that our target is to categorize a video into one of objects in the dataset, and the only label we have for training the model is the video-level label – the object class of each video. Inspired by the objectness methods [1, 3, 5, 14, 25], which reduce the search space and generate the potential object areas in an image in unsupervised ways, and Latent SVM methods [10, 28], which solve the weakly supervised problems by introducing latent variables, we propose a novel algorithm for object recognition in video, which only uses the weak video-level label, jointly inferring the object location in each frame and recognizing the object in the video. Once the object location is obtained, the background information is added as a complement of the object information to improve the discriminability, and temporal information is incorporated to keep the consistency of recognition decision in a video.

To demonstrate the benefits of the proposed method, we collected a video dataset for object recognition from YouTube, named *YouTube* dataset. 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. To the best

of our knowledge, there is no such dataset publicly available for video-based object recognition. The introduction of the dataset is another contribution of this work.

2. Related work

The majority of algorithms for object recognition [4, 17, 18, 26] assume the input is a single image. In recent years, there has been an increasing attention to object recognition in videos [12, 15, 16, 21, 23, 24]. There are mainly two directions among the previous work. One is to utilize the temporal information in video and improve local video descriptors by feature tracking [12, 15, 23, 24]. The other is to first run the image-based recognition on individual frames of a video and subsequently voting for the video class [16, 21]. Both ways have been proven to be successful.

This paper follows the voting method, aiming at solving the recognition problem by localizing the object in each frame of a video. Since only video-level labels are provided, the localization and recognition problem is in a weakly supervised scenario. There are several previous works on weakly supervised object localization [8, 13, 19], but they mainly focus on the binary localization problem, rather than utilize it for the multi-class object recognition one. These works mainly rely on the objectness methods to initialize the localizer, reducing the search space and generating the location candidates, such as selective search [25], objectness [1], efficient subwindow search [14], large-displacement optical flow [5], and superpixels merging [3]. We use the selective search method [25] for initialization in this paper.

For weakly supervised problem, Multiple Instance Learning (MIL) [2, 11] and Latent SVM [10, 28] are the two main learning frameworks for it. MIL has also been used for image-based object recognition and localization [7, 29]. It can be naturally extended to solve the problem of video-based object recognition, by treating a video as a bag of images, a video of the target object as a positive bag, and a video without the target object as a negative bag. Latent SVM is more popular in solving the weakly supervised problem, including weakly supervised action localization [22], view selection for action recognition [27], etc.

In this paper, we follow the Latent SVM framework to jointly solve the object recognition and localization problems in videos. Different from the previous localization work, our final target is the multi-class object recognition in video, while localization is only the intermediate product to help the object recognition.

3. Video-based object recognition

3.1. Model formulation

Given only the object class label of training videos, our goal is to jointly recognize the object class of a testing video and infer the location of the target object in each frame. Potential candidate locations in each frame are generated by a selective search method [25] and represented by a latent variable h.

Given a video $\mathbf{x}$, we select uniformly sampled sets of frames $\{\mathbf{x}_t | t = 1, 2, ...\}$ from $\mathbf{x}$. Similar to the Latent SVM [10, 28], we set $f_w(\mathbf{x}_t, y)$ as the score function for a video frame $\mathbf{x}_t$ and a class label y, measuring the compatibility between the frame $\mathbf{x}_t$ and the label y. $f_w(\mathbf{x}_t, y)$ is a linear function consisting of three components:

$$f_{\mathbf{w}}(\mathbf{x}_t, y) = f_{\mathbf{w}_O, \mathbf{w}_B, \mathbf{w}_T}(\mathbf{x}_t, y) = \max_{h_t} [\mathbf{w}'_O \psi(\mathbf{x}_t, y, h_t) + \mathbf{w}'_B \psi(\mathbf{x}_t, y, \bar{h_t}) + \mathbf{w}_T \psi(\mathbf{x}_t, h_t; \mathbf{x}_{t-1}, h^*_{t-1})] \quad (1)$$

where latent variable h_t represents the pixels of the image at frame $\mathbf{x}_t$ that belong to the object and $\bar{h_t}$ the rest. In other words, h_t denotes the "foreground" and $\bar{h_t}$ the "background". $\psi(\mathbf{x}_t, y, h_t)$ is the joint feature mapping of the video frame $\mathbf{x}_t$, the label y and the object location h_t. $\mathbf{w}$ is the parameter of the score function that needs to be learned. The model is composed of three components: object potential $\mathbf{w}'_O \psi(\mathbf{x}_t, y, h_t)$, background potential $\mathbf{w}'_B \psi(\mathbf{x}_t, y, \bar{h_t})$, and temporal potential $\mathbf{w}_T \psi(\mathbf{x}_t, h_t; \mathbf{x}_{t-1}, h^*_{t-1})$. The model parameter $\mathbf{w}$ is the concatenation of the three components $\{\mathbf{w}_O; \mathbf{w}_B; \mathbf{w}_T\}$.

Object Potential $\mathbf{w}'_O \psi(\mathbf{x}_t, y, h_t)$: It measures the compatibility of the video frame $\mathbf{x}_t$, the label y and the object location h_t. The joint feature vector $\psi(\mathbf{x}_t, y, h_t)$ aggregates the feature points in the object area h_t. It helps to a) localize the object (compact bounding box), b) remove the background clutter and c) get a cleaner representation for the object.

Background Potential $\mathbf{w}'_B \psi(\mathbf{x}_t, y, \bar{h_t})$: Background information $\bar{h_t}$ is an important complement of object information since it adds discriminability to recognition, i.e. airplanes flying in sky, cars running on road, ships sailing in water. The joint feature vector $\psi(\mathbf{x}_t, y, \bar{h_t})$ aggregates the feature points in the background area $\bar{h_t}$, while preventing features on the object area leaking into the background, improving, thus, the localization accuracy.

Temporal Potential $\mathbf{w}_T \psi(\mathbf{x}_t, h_t; \mathbf{x}_{t-1}, h^*_{t-1})$: It enforces a tracking constraint, i.e. at the t^{th} frame ($\mathbf{x}_t$) the object location h_t should be similar with the one obtained previously at $(t-1)^{th}$ frame ($\mathbf{x}_{t-1}$).

$$\psi(\mathbf{x}_t, h_t; \mathbf{x}_{t-1}, h^*_{t-1}) = ||\psi(\mathbf{x}_t, h_t) - \psi(\mathbf{x}_{t-1}, h^*_{t-1})||_1 \quad (2)$$

where h^*_{t-1} is the inferred object location at the $\mathbf{x}_{t-1}$ frame. $\psi(\mathbf{x}_{t-1}, h^*_{t-1})$ and $\psi(\mathbf{x}_t, h_t)$ denote vectors of feature points in the extracted object bounding box at frames $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, respectively. $\mathbf{w}_T$ is a scaler value that controls the weight of the temporal potential. We account for the temporal information in a video to keep the consistency of localization and increase the recognition performance.

Figure 1. Representative frames of *Youtube* dataset.

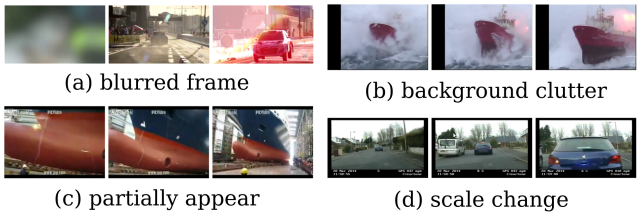

Figure 2. Examples of challenging frames in *YouTube* dataset.

3.2. Learning and inference

Let $\{\mathbf{x}^1, ..., \mathbf{x}^n\}$ denote a set of training videos and y^i the class label of object in video $\mathbf{x}^i$. The goal is to learn the model parameter $\mathbf{w} = \{\mathbf{w}_O; \mathbf{w}_B; \mathbf{w}_T\}$, which facilitates the efficient classification of a testing video into the correct class along with the accurate localization of the target object at each frame. Let $\{\mathbf{x}_t^i\}$ be the set of frames of video $\mathbf{x}^i$, uniformly sampled from the video, and $y_t^i = y^i$ the corresponding class label of each frame. We utilize the latent SVM framework [10, 28] for learning, i.e.:

$$\underset{\mathbf{w}_O,\mathbf{w}_B,\mathbf{w}_T,\sigma_t^i}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}_O||^2 + \frac{1}{2}||\mathbf{w}_B||^2 + \frac{1}{2}||\mathbf{w}_T||^2 + C\sum_{i=1}^{n}\sum_{t}\sigma_t^i \tag{3}$$

$$\text{s.t. } f_{\mathbf{w}_O,\mathbf{w}_B,\mathbf{w}_T}(\mathbf{x}_t^i, y_t^i) - f_{\mathbf{w}_O,\mathbf{w}_B,\mathbf{w}_T}(\mathbf{x}_t^i, y) \geq 1 - \sigma_t^i \quad \forall i, \ \forall t, \ \forall y \neq y_t^i,$$

$$\sigma_t^i \geq 0 \quad \forall i, \ \forall t.$$

The constrained optimization in Eq. (3) is equivalent to the following unconstrained problem:

$$\underset{\mathbf{w}_O,\mathbf{w}_B,\mathbf{w}_T,\sigma_t^i}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}_O||^2 + \frac{1}{2}||\mathbf{w}_B||^2 + \frac{1}{2}||\mathbf{w}_T||^2 + C\sum_{i=1}^{n}\sum_{t}\max\{0, 1 - f_{\mathbf{w}_O,\mathbf{w}_B,\mathbf{w}_T}(\mathbf{x}_t^i, y_t^i) + f_{\mathbf{w}_O,\mathbf{w}_B,\mathbf{w}_T}(\mathbf{x}_t^i, y)\} \tag{4}$$

We use the Non-convex Regularized Bundle Method (NRBM) [9] to optimize Eq. (4). NRBM combines bundle methods and cutting plane techniques, while it 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 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 ϵ.

Given the model parameter $\mathbf{w}$, we want to jointly predict the object class and estimate its position in a new testing video $\mathbf{x}$. The correct class and the bounding box parameters for the frame t are the ones that maximize the score:

$$(y_t^*, h_t^*) = \underset{y_t,h_t}{\text{argmax}} \ f_{\mathbf{w}_O,\mathbf{w}_B,\mathbf{w}_T}(\mathbf{x}_t, y_t) \tag{5}$$

Finally, the object class of the video $\mathbf{x}$ is obtained through voting the recognition results of all frames.

4. Experiments

4.1. Dataset

For evaluation, we collected a video dataset for object recognition from YouTube. It contains 300 videos of 10 object categories: airplane, ring, ship, car, guitar, pipe, horse, teapot, pram, bicycle. The representative frames of each category are shown in Figure 1. Each category contains 30 videos of different object instances of this category, from which 10 videos are randomly chosen for training and the other 20 videos are for testing. Each video lasts for 10 to 15 seconds, and 3 frames per second are uniformly sampled for recognition.

The *YouTube* dataset is very challenging. These in-the-wild videos from the Internet are totally unconstrained, with different style, quality, and resolution. Some videos are taken by mobile phones hold in hands, leading to blurred frames (Figure 2(a)). The videos are mainly taken from the real-world wild environment, with all kinds of illumination conditions (Figure 2(a)) and cluttering background (Figure 2(b)). The objects in one category are also very diverse. For example, the ships in ship category vary from ancient boat to modern warship, shown in Figure 3. In teapot category, there are 42 different teapots in the 30 videos, meaning that some videos contain multiple teapots. The appearance of an object in a video also changes largely: sometimes it is as small as a dot; sometimes it is so big that it is only partially shown in the video, as shown in Figure 2(c)(d).

Note that the existing *YouTube-Objects* dataset [20] is for object detection task. It only contains around ten to twenty videos for each class, meaning that there are not enough videos for our video-based object recognition task.

4.2. Experiment results and discussion

We have compared the proposed method with the Accumulating Frame (Accum-Frame) [16, 21] and Multiple

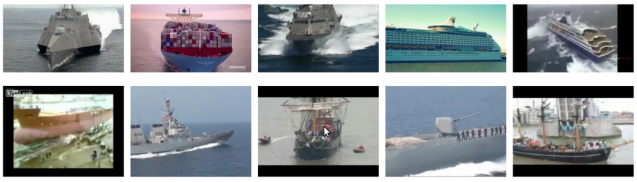

Figure 3. Examples of different ships in ship category.

Figure 4. Examples of localization result: green – good, yellow – ok, red – bad. Best viewed in color.

Table 1. Comparison experiment results on *YouTube* dataset. Our proposed method achieves the best accuracy.

Method	Accum-Frame	MIL	Our
Accuracy	57.5%	63.5%	**69.5%**

Table 2. Comparison with variant methods. OB: Object; BG: Background; TP: Temporal information.

Method	OB	OB+BG	OB+BG+TP
Image-based	54.34%	57.84%	59.28%
Video-based	63%	67.5%	**69.5%**

Instance Learning (MIL) [2, 11] methods for video-based object recognition.

We use Dense SIFT (DSIFT) [4, 18] feature for all methods in the experiment. Then Pyramid Histogram Of visual Words (PHOW) [4] and Bag-of-Words (BoW) [23] were used for aggregating the DSIFT descriptors. Two layer PHOW was used as the object representation in the proposed method and the image representation in the compared methods. BoW was used as the background representation in the proposed method. The codebook size for both PHOW and BoW were 512. Other settings follow the standard way as in [4] and [23].

In Accum-Frame method, we first conduct the frame-based recognition. 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 [6] with one-vs-one setting. Then, three voting methods – Hard, Soft, and KNN voting – are used to accumulate the results of frame-based recognition to get the video class. They respectively use the label results, the probability estimation of SVM, and K (empirically set as 20) best frames to vote for the video class.

For MIL method, since a video can be seen as a bag of frames (instances), MIL can be naturally used to handle the noisiness of video data and perform the video-based object recognition. The video of the target object is treated as a positive bag, in which the frames containing the target object are seen as the positive instances, while other frames which do not contain the target object or are corrupted hence very noisy, are deemed as the negative instances. The videos of other classes, which do not contain the target object, are treated as negative bags. The MILSVM algorithm [11] is used here.

Table 1 shows the experiment results on the *YouTube* dataset. All parameters were set to report the best accuracies. For the proposed method, the model slack variable coefficient was set as $C = 10^{-4}$, and the optimization stopping criterion was $\epsilon = 0.01$. The result of Accum-Frame was obtained from KNN voting, which boosted the image-based recognition accuracy 49.16% to 57.5%. The low accuracy of image recognition also demonstrates the challenges of the *YouTube* dataset. The Hard and Soft voting methods got 55.5% and 57% respectively, slightly worse than KNN voting. All of these results are significantly better than image-based recognition, indicating the importance of accumulating information in a video for object recognition. MIL got a better accuracy 63.5% than Accum-Frame. This is because MIL can iteratively filter out the noisy frames (negative instances) and select the good frames (positive instances) in a training video (a positive bag) to train the classifier, yielding a more accurate model than Accum-Frame. Although KNN voting in Accum-Frame can also select K good frames to vote for the video class, this selection process is only in testing, which is not able to filter out the noise information during training and leads to a not very accurate model. Our proposed method achieves the best accuracy 69.5%. The weakly supervised localization of our method helps to separate the object and the background, so that the feature points on the object and the background are aggregated into two vectors separately. It not only makes the object representation cleaner, but also adds the background information to improve the accuracy. The temporal information of the video is also utilized to further improve the result. In contrast, without the bounding box annotation, the compared methods simply aggregate all the features in a image into one feature vector, which is less discriminative.

Table 2 reports the performance of two variants of the proposed method. OB is the method which only use the object potential in the model. With the help of weakly supervised localization, it has already achieved comparable result 63% with MIL. In OB+BG, background potential is added, which largely improve the recognition accuracy to 67.5%. It indicates that the background also has certain discriminability and is a good complement to the object itself. OB+BG+TP is the proposed method, with the temporal potential added in the model. As the particular information in video, temporal information helps keep the consistency of the recognition decision in a video, and finally boosts the

accuracy to 69.5%. The video-based results are obtained by voting technique on corresponding image-based results.

Figure 4 shows some localization results in testing. Most of them are good, with compact bounding boxes on the exact target objects. Some of them are not perfect, with loose bounding boxes on parts of the target objects. Some are very bad due to the noise, such as the clutters.

The proposed method is efficient. In training, it took about 5 hours with the maximum iteration of 300. In testing, it took only 0.19 second to classify a video on average. This excludes the time for object candidates generation and feature extraction, which are decided by their own algorithms. These timing figures were measured on an Intel Core i7 3.4GHZ×8 processor with 8GB RAM, for a Matlab implementation.

5. Conclusion

We proposed a novel algorithm based on Latent SVM and a new *Youtube* dataset for video-based object recognition. With only the video-level label in training, the proposed algorithm is able to jointly recognize the object in a testing video and infer its location in each video frame. It comprehensively utilizes the background and temporal information of objects in video to accomplish the task. The experiments on the *Youtube* dataset demonstrate the effectiveness of the proposed method. In future, we plan to combine Latent SVM and MIL to jointly perform the object localization and view selection in video for object recognition.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence*, 2012. 1, 2

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002. 2, 4

[3] M. V. D. Bergh, G. Roig, X. Boix, S. Manen, and L. V. Gool. Online video seeds for temporal window objectness. In *ICCV*, 2013. 1, 2

[4] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *ICCV*, 2007. 1, 2, 4

[5] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *CVPR*, 2009. 1, 2

[6] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011. 4

[7] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014. 2

[8] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010. 2

[9] T.-M.-T. Do and T. Artières. Large margin training for hidden Markov models with partially observed states. In *ICML*, 2009. 3

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence*, 2010. 1, 2, 3

[11] P. V. Gehler and O. Chapelle. Deterministic annealing for multiple-instance learning. In *International conference on artificial intelligence and statistics*, 2007. 2, 4

[12] V. Gouet-Brunet and B. Lameyre. Object recognition and segmentation in videos by connecting heterogeneous visual features. *Computer Vision and Image Understanding*, 2008. 2

[13] G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *NIPS*, 2009. 2

[14] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008. 1, 2

[15] T. Lee and S. Soatto. Video-based descriptors for object recognition. *Image and Vision Computing*, 2011. 2

[16] B. Li, R. Chellappa, Q. Zheng, and S. Z. Ser. Model-based temporal object verification using video. *IEEE Tran. on Image Processing*, 2001. 2, 3

[17] F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 1, 2

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 1, 2, 4

[19] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011. 2

[20] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 3

[21] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010. 2, 3

[22] 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. 2

[23] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2, 4

[24] D. Stavens and S. Thrun. Unsupervised learning of invariant features using video. *In CVPR*, 2010. 2

[25] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011. 1, 2

[26] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 1, 2

[27] X. Wu and Y. Jia. View-invariant action recognition using latent kernelized structural svm. In *ECCV*, 2012. 2

[28] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, 2009. 1, 2, 3

[29] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *NIPS*, 2005. 2