Empowering Semantic Indexing with Focus of Attention

Kimiaki Shirahama*, Tadashi Matsumura†, Marcin Grzegorzek*, and Kuniaki Uehara†

* Pattern Recognition Group, University of Siegen
† Graduate School of System Informatics, Kobe University
Email: kimiaki.shirahama@uni-siegen.de, tadashi@ai.cs.kobe-u.ac.jp,
marcin.grzegorzek@uni-siegen.de, uehara@kobe-u.ac.jp

***Abstract*—This paper addresses Semantic INdexing (SIN) to detect concepts like *Person* and *Car* in video shots. One main obstacle is the abundant information contained in a shot where multiple concepts are displayed at the same time. In other words, the detection of a target concept is adversely affected by other concepts which are incidentally shown in the same shot. We assume that a user can recognise the target concept when it appears in a salient region which attracts his/her attention. Based on this, we introduce a SIN method which utilises Focus of Attention (FoA) to extract a salient region in a shot, and constructs a feature emphasising that salient region. In addition, we develop a Weakly Supervised Learning (WSL) method to efficiently create training shots for FoA, and a shot filtering method to examine the usefulness of salient regions. Experimental results show the effectiveness of our SIN method using FoA.**

***Keywords*–Semantic indexing; Focus of attention; Weakly supervised learning; Usefulness of salient regions.**

I. INTRODUCTION

For effective filtering, categorisation, browsing and retrieval of large-scale video data, one key technology is 'Semantic INdexing' (SIN) to detect human-perceivable concepts (e.g., *Person*, *Car* and *Building*) in shots [1]. SIN is formulated as a binary classification problem where shots displaying a target concept are distinguished from the rest of the shots. The currently most popular approach is to represent a shot as a collection of *descriptors* which characterise patches (small regions) in the shot [2][3]. To make the following discussion clear, we define a descriptor as the representation of a patch, and a feature as the representation of a shot based on a set of descriptors. The effectiveness of this feature is attributed to considering many descriptors extracted from various patches. Even if the target concept is partially invisible due to the occlusion by other concepts or the camera setting, the feature includes descriptors extracted from patches corresponding to the visible part of the target. However, many concepts other than the target are displayed in a shot. For example, the top-left shot in Figure 1 includes the target concept *Car* and many others like *Building*, *Road* and *Sky*. Nonetheless, most of the existing methods [2][3] do not consider whether each patch belongs to the target concept or not. The resulting feature is affected by patches of other concepts, and the detection performance of the target concept is degraded.

To effectively spotlight a target concept, we propose a SIN method using 'Focus of Attention' (FoA). FoA implements selective attention that is a brain mechanism to determine which region in a video frame (or image) attracts the user [4]. Such an attractive region is called a *salient region*. FoA is beneficial to develop a system which can sort out visual information according to human perception. In our case, we assume that the target concept can be recognised by a user when it appears in salient regions. In other words, the user is unlikely to realise appearances of the target concept in non-salient regions. These appearances are trivial and useless for subsequent processes like video categorisation, browsing and retrieval. Therefore, we use FoA to increase priorities of salient regions and decrease those of non-salient regions. This prioritisation enables us to construct a feature which emphasises an appearance of the target concept, so that its detection performance can be improved.

FoA consists of two main processes, *bottom-up* and *top-down*. The former implements human attention driven by stimuli acquired from the external environment, where salient regions are detected based only on features. However, these salient regions are not so accurate because of the *semantic gap*, which is the lack of agreement between automatically extractable features and human-perceived semantics [5]. Thus, the top-down process implements attention which is driven by prior knowledge and expectation in the internal human mind. This is typically formulated in the framework of machine learning, where salient regions in test shots are detected by referring to training shots in which salient regions are annotated in advance. More concretely, inaccurate salient regions obtained by the bottom-up process are refined based on salient regions in training shots.

To incorporate FoA into SIN, we address the following two issues: The first is that salient regions significantly vary depending on camera techniques and shooting environments. A large number of training shots is needed to accurately detect diverse salient regions. However, due to a tremendous number of video frames in shots, it requires prohibitive cost to manually prepare many training shots. Thus, we develop an FoA method using 'Weakly Supervised Learning' (WSL) where a classifier to predict precise labels is constructed only using loosely labelled training data [6]. In our case, this kind of training data are shots that are annotated only with the presence or absence of a target concept. Using these training shots, we build a classifier which can identify the region of the target concept in a shot. In the top-down process, regions of the target concept in training shots are identified by the classifier and regarded as annotated salient regions.

The second issue is the discrepancy that salient regions do not necessarily contain a target concept. The reason is twofold: Firstly, there is the difficulty of objectively judging whether the target concept appears in a salient region or not. In other words, we can only use training shots where the presence of the target concept is annotated without considering its saliency.

 ISBN: 978-1-61208-400-8

For example, in Figure 1, training shots annotated with the presence of the target concept *Car* include the bottom-right shot where the car is shown in the small background region. It is impossible or unreasonable to regard this region as salient. The second reason for the discrepancy is possibly occurring errors in FoA. Even if the region of the target concept is salient for humans, another region may be falsely regarded as salient. A feature based on such a salient region which is discrepant with the region of the target concept incorrectly emphasises a non-target concept. To alleviate this, we develop a method which filters out shots where the target concept is unlikely to appear in salient regions, using regions predicted by the classifier in WSL. This enables SIN integrated with FoA to appropriately capture a characteristic feature for the target concept appearing in salient regions.

This paper is organised as follows: In Section II, we describe the novelties of our method by discussing insufficiencies of existing FoA methods with regard to SIN. Section III presents our method by sequentially explaining the FoA (consisting of the bottom-up and top-down processes), WSL and SIN modules. In Section IV, we evaluate our method using large-scale video data. Section V concludes this paper by providing a future extension of our method.

II. Related Work

Existing machine learning approaches for the top-down process are typically based on the *contextual cueing* which means "a user can easily search a particular object among many objects, if he/she saw the same or similar spatial layout of objects in the past". These approaches construct a model which properly integrates features extracted in the bottom-up process, using recorded eye-fixations or labelled salient regions as training data. Itti and Koch proposed an approach to compute the optimal weights to integrate features [7]. Kienzle *et al.* proposed a non-parametric approach to build a model using recorded eye-fixations [8]. Furthermore, Li *et al.* proposed an approach which simultaneously constructs a set of models to integrate features using multi-task learning [9]. However, these approaches assume the availability of training videos where salient regions are labelled. Compared to this, we incorporate WSL into FoA so as to only require training shots which are annotated just with the presence or absence of a target concept. In addition, the approaches described above are only evaluated on shots which necessarily contain some salient regions. In other words, they do not consider what is shown in a detected salient region, or how to use it in a subsequent application. This paper explores how to utilise salient regions for SIN and presents a method for filtering shots where salient regions show non-target concepts.

III. Concept Detection Using FoA

Figure 1 illustrates an overview of our SIN method where the target concept is *Car*. We call training shots annotated with the presence and absence of the target concept *positive shots* and *negative shots*, respectively. The bold arrows in Figure 1 show the dominant flow where the FoA module creates a *saliency map* for each training shot. This map is an image which represents the saliency of each pixel. Figure 1 shows saliency maps obtained for the positive and negative shots at the top. As pixels have higher saliencies, they are depicted as brighter. The positive and negative shots in Figure 1 are appropriately associated with salient regions where *Car*

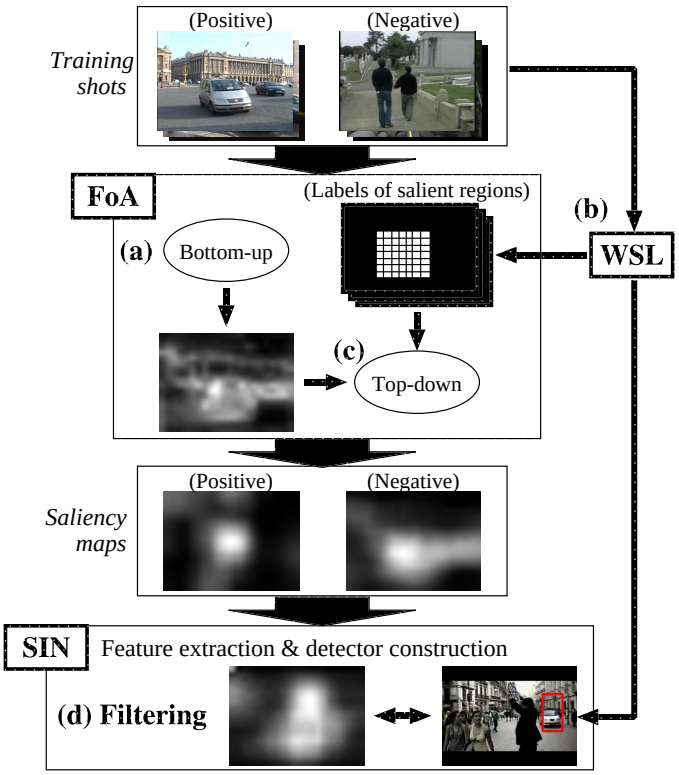

Figure 1. An overview of our SIN method using FoA with WSL.

and *Person* are shown, respectively. The SIN module uses saliency maps to consider the saliency of each patch from which a descriptor is extracted. Thereby, the feature for each training shot is constructed by weighting descriptors based on saliencies of their patches. Then, a *detector* is constructed to identify the target concept in test shots.

The FoA module works as follows: First, the bottom-up process in Figure 1 (a) computes a saliency map for each training shot (for short, 'bottom-up saliency map'). Meanwhile, WSL in Figure 1 (b) builds a classifier using training shots, in order to identify regions of the target concept in positive shots. By regarding these regions as salient, the top-down process is performed as depicted in Figure 1 (c). This refines the bottom-up saliency map for each training shot into the final saliency map. Also, the filtering process in Figure 1 (d) examines whether each positive shot should be used to build a detector, by comparing its saliency map to the region detected by WSL (red rectangle). If the region is not salient, the positive shot is excluded from the training shots. This is because such positive shots mislead the detector to favour non-target concepts. Below, we describe the bottom-up and top-down processes, WSL method, and SIN method.

Bottom-up process: This simulates a retina model of a human to detect salient regions as the ones where features are different from those of surrounding regions [9]. For each video frame in a shot, the bottom-up process creates six 'feature maps', each of which describes every pixel using a different feature, that is, luminance, red-green contrast, blue-yellow contrast, flicker, motion direction, or motion strength [9]. Then, to simulate the mechanism of the horizontal cells of a retina, wavelet transform is performed per feature map so that three 'wavelet images' with different resolutions are created. Afterwards, the mechanism of the bipolar cells is simulated where high-pass

 ISBN: 978-1-61208-400-8

filters are used to extract high-frequency components in three directions for each wavelet image. As a result, three 'edge images' are created where the difference of a region from surroundings is represented as edges. For noise reduction and computational efficiency, each edge image is smoothed by a Gaussian filter and divided into 18×22 macro-blocks, where each block is represented as the average of included pixel values. Finally, a bottom-up saliency map is created by averaging values of the same macro-block in all of 54 edge images (6 feature maps $\times$ 3 wavelet images $\times$ 3 edge images). The saliency of each macro-block is represented by a real number between 0 and 1 (see [9] for more detail).

Top-down process: Assuming that salient regions in positive shots are labelled by WSL described below, the top-down process performs *multi-task learning* which effectively solves multiple 'related tasks' at the same time by extracting the information shared among them [9]. We define a task as the refinement of the bottom-up saliency map of each positive shot. This task is related to tasks for other positive shots in the sense that they are taken by similar camera techniques and in similar shooting environments. Thus, we refine their bottom-up saliency maps using the same set of functions, which individually represent a different linear combination of features used in the bottom-up process. The bottom-up saliency map of each positive shot is refined by the weighted fusion of these functions' outputs, so that the refined saliency map matches the labelled salient region. We use an EM-like algorithm to optimise functions, and weights of their outputs for each positive shot [9].

Based on the contextual cueing described in Section II, the top-down process for a test shot begins with selecting the positive shot which has the most similar spatial layout to that of the test shot. The weighted fusion used for this positive shot is re-used to refine the bottom-up saliency map of the test shot. The similarity between two shots in terms of spatial layouts is computed as their cosine similarity based on features used in the bottom-up process.

Weakly Supervised Learning: Given training shots annotated only with the presence or absence of a target concept, we construct a classifier which can identify its region in a shot. This WSL is achieved so that the classifier characterises regions which are contained in positive shots, but are not contained in negative shots [6]. The target concept is considered to appear in these regions. The classifier is optimised by iterating the following two processes: The first examines regions in each training shot to find the 'best region' which maximises the output of the current classifier. The other process updates the classifier using the newly found best regions. As a result, the classifier outputs high values for the best regions in positive shots, while low values are assigned to all regions including the best ones in negative shots. The best regions in positive shots are regarded as the labelled salient regions.

Since a video frame in a shot contains a huge number of possible regions, the aforementioned WSL needs to efficiently find the best region. To this end, we implement the classifier as a linear SVM based on quantised SIFT descriptors, called 'Visual Words' (VWs) [10]. First, we extract SIFT descriptors from patches which have the radius of 10 pixels and are located at every sixth pixel in each video frame. Then, randomly sampled one million SIFT descriptors are grouped into $1,000$ clusters where each cluster centre is a VW. Every SIFT descriptor is quantised into the most similar VW. Based on this, a region is represented by a feature (histogram) where each dimension represents the frequency of a VW in this region. In particular, for any region, the output of the linear SVM can be computed by simply counting the frequency of each VW. This enables us to estimate the 'upper bound' for a set of regions [10]. Here, no region in the set takes the output larger than the upper bound. The best region can be efficiently found by discarding many sets of regions for which upper bounds are small. With this efficient search, we can refine the classifier (linear SVM) through many iterations.

Semantic Indexing: Given a target concept, we create the feature of each training shot as a histogram where each bin represents the 'weighted' frequency of a VW. We check the saliency map to obtain the saliency of each patch from which a SIFT descriptor is extracted. This saliency is used to weight the frequency of the VW associated with the SIFT descriptor. As a result, the feature emphasises frequencies of VWs in the salient region where the target concept probably appears. Using such features, a detector is constructed as a non-linear SVM with Radial Basis Function (RBF) kernel.

Before constructing the detector, we use the classifier in WSL to examine whether salient regions in positive shots include the target concept or not. We assume that the target concept is salient if its region is large. Hence, we filter out a positive shot if the best region is very small, or the classifier's output for this region is very small. This filtering is also applied to test shots. For test shots where salient regions fail to cover the target concept, features obtained by weighting VWs' frequencies undesirably emphasise non-target concepts. To avoid this, the filtering aims to distinguish test shots where salient regions certainly include the target concept from the others. For the former, we extract features by weighting VWs' frequencies, while features for the latter are extracted without weighting. Finally, the list of sorted test shots in terms of outputs by the detector is returned as the SIN result.

IV. Experimental Results

We first examine the effectiveness of FoA using WSL by targeting three concepts *Person*, *Car* and *Explosion_Fire*. For each concept, we use $1,000$ positive shots and $5,000$ negative shots in TRECVID 2009 video data [1]. The performance is evaluated on $1,000$ test shots where the ground truth of salient regions is manually provided. We compare two FoA methods, *WSL* and *Manual*, which use positive shots where salient regions are labelled by WSL and by the manual method, respectively. Figure 2 shows ROC curves for *WSL* and *Manual*. Each curve is created by calculating true positive (TP) and false positive (FP) rates using different thresholds. Here, a pixel in a saliency map is regarded as salient if its saliency is larger than a threshold. A TP is the number of pixels which are correctly detected as salient, and a FP is the number of pixels falsely detected as salient. Figure 2 shows that, for all concepts, ROC curves of *WSL* and *Manual* are nearly the same. This means that FoA can be appropriately performed even using salient regions labelled by WSL.

As another evaluation measure, an AUC represents the area under an ROC curve. A larger AUC indicates superior performance where a high TP is achieved for a small FP. Figure 2 presents that *WSL*'s AUCs are nearly the same or even larger than those of *Manual*. Note that several regions where a target concept does not appear are falsely detected by WSL, and used as labelled salient regions in the top-down

 ISBN: 978-1-61208-400-8

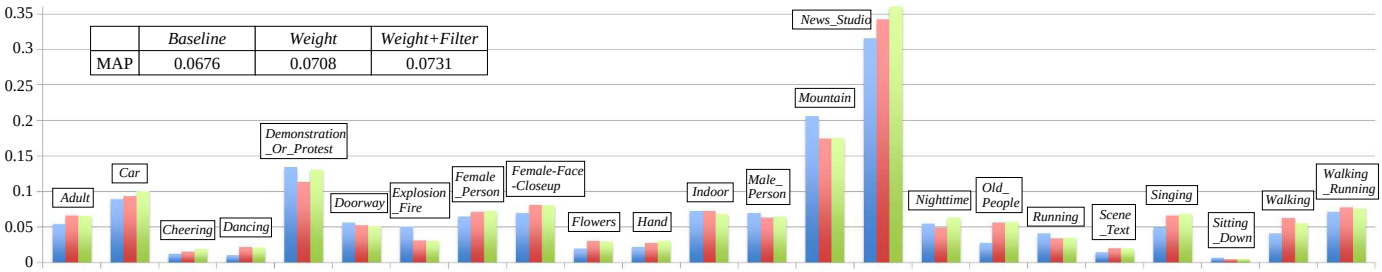

Figure 3. Performance comparison among *Baseline*, *Weight* and *Weight+Filter*.

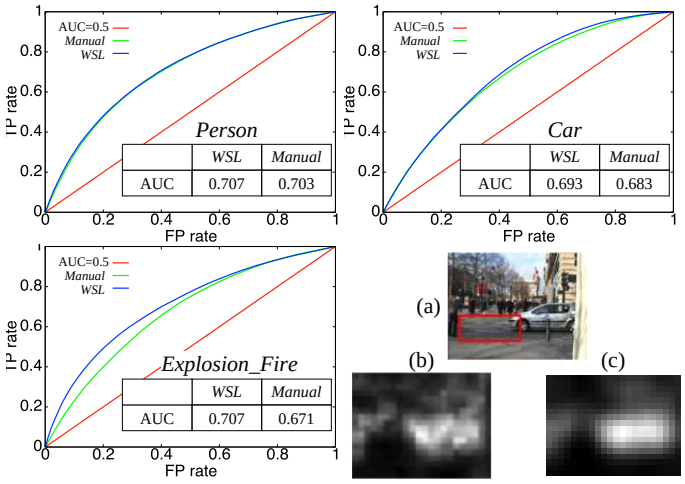

Figure 2. Performance comparison between *WSL* and *Manual*.

process. For example, at the bottom-right of Figure 2, the red rectangular region in the image (a) is falsely regarded as showing a car. However, as seen from the bottom-up saliency map marked by (b), the saliency of this region is very low, so it cannot be a salient region even with the refinement by the top-down process (see the image (c)). Like this, errors in WSL are alleviated based on saliencies obtained by the bottom-up process. In other words, FoA works appropriately as long as regions obtained by WSL are mostly correct.

Next, we evaluate the performance of our SIN method using FoA. According to the official instruction of TRECVID 2011 SIN light task [1], we select 23 target concepts shown in Figure 3. For each target, a detector is constructed with $30,000$ training shots collected from $240,918$ shots in $11,485$ development videos, and tested on $125,880$ shots in $8,215$ test videos. To examine the effectiveness of weighting descriptors based on FoA and that of shot filtering, we compare three methods *Baseline*, *Weight* and *Weight+Filter*. *Baseline* and *Weight* use features defined by original frequencies and weighted frequencies of VWs, respectively. *Weight+Filter* extends *Weight* by adding the shot filtering process.

Figure 3 shows the performance comparison in form of a bar graph. For each concept, the left, centre and right bars represent Average Precisions (APs) of *Baseline*, *Weight* and *Weight+Filter*, respectively. A larger AP indicates a higher performance. For each method, we also exhibit the Mean AP (MAP) which is the mean of APs over 23 concepts. Figure 3 illustrates that *Weight* outperforms *Baseline* for many concepts. The MAP of the former is about 5% higher than that of the latter. This validates the effectiveness of FoA for SIN. In addition, *Weight+Filter*'s MAP indicates that adding shot filtering improves *Weight*'s MAP by about 3%. This verifies the effectiveness of shot filtering.

V. Conclusion and Future Work

In this paper, we introduced a SIN method using FoA with WSL. Experimental results showed both the validity of incorporating WSL into FoA and the effectiveness of FoA for SIN. Figure 3 shows that FoA causes the performance degradation for some concepts such as *Explosion_Fire* and *Mountain*. One main reason is non-rectangular shapes of these concepts, while our WSL method can only identify rectangular regions. In other words, rectangular regions are too coarse to precisely localise non-rectangular concepts, and inevitably include other concepts. As a result, the top-down process does not work well. Hence, we will extend our WSL method by adopting the efficient search algorithm for regions with arbitrary shapes [11].

Acknowledgment

This work was carried out when Tadashi Matsumura belonged to Graduate School of System Informatics, Kobe University.

References

[1] A. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in Proc. of MIR 2006, 2006, pp. 321–330.

[2] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, 2010, pp. 1582–1596.

[3] N. Inoue and K. Shinoda, "A fast and accurate video semantic-indexing system using fast map adaptation and gmm supervectors," IEEE Trans. Multimed., vol. 14, no. 4, 2012, pp. 1196–1205.

[4] S. Frintrop, E. Rome, and H. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," ACM Trans. Appl. Percept., vol. 7, no. 1, 2010, pp. 6:1–6:39.

[5] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, 2000, pp. 1349–1380.

[6] M. Nguyen, L. Torresani, F. De la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: a joint learning process," in Proc. of ICCV 2009, 2009, pp. 1925–1932.

[7] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," J. Electron. Imaging, vol. 10, no. 1, 2001, pp. 161–169.

[8] W. Kienzle, F. Wichmann, B. Schoelkopf, and M. Franz, "A nonparametric approach to bottom-up visual saliency," in Proc. of NIPS 2006, 2006, pp. 689–696.

[9] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," Int. J. Comput. Vis., vol. 90, no. 2, 2010, pp. 150–165.

[10] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in Proc. CVPR 2008, 2008, pp. 1–8.

[11] S. Vijayanarasimhan and K. Grauman, "Efficient region search for object detection," in Proc. of CVPR 2011, 2011, pp. 1401–1408.

 ISBN: 978-1-61208-400-8