# Spatio-Temporal Saliency for Action Similarity

G.J. Burghouts, S.P. van den Broek, R.J.M. ten Hove
TNO, Intelligent Imaging
www.tno.nl/IntelligentImaging
gertjan.burghouts@tno.nl

## Abstract

*Human actions are spatio-temporal patterns. A popular representation is to describe the action by features at interest points. Because the interest point detection and feature description are generic processes, they are not tuned to discriminate one particular action from the other. In this paper we propose a saliency measure for each individual feature to improve its distinctiveness for a particular action. We propose a spatio-temporal saliency map, for a bag of features, that is specific to the current video and to the action of interest. The novelty is that the saliency map is derived directly from the SVM's support vectors. For the retrieval of 48 human actions from the visint.org database of 3,480 videos, we demonstrate a systematic improvement across the board of 35.3% on average and significant improvements for 25 actions. We learn that the improvements are achieved in particular for complex human actions such as giving, receiving, burying and replacing an item.*

## 1. Introduction

To find corresponding human actions in realistic videos a system needs the following key elements: robustness / invariance of the features to changing recording conditions, sensitivity to the motion patterns and appearance, selectivity of the feature representation for the current action, and good discrimination between the positives and negatives by a robust classifier [1-4]. Simple bag-of-features action detectors e.g. [5,6], and more advanced extensions that exploit spatio-temporal layout [7], feature fusion [8], world-knowledge [9] and dealing with uncertainty [10] have demonstrated to be very effective for the task of action detection, including quite complex actions such as digging in the ground, falling onto the ground, and chasing somebody. Yet, for the detection of more complex actions, such as the exchange of an item, or burying or hauling something, the standard bag-of-features action detectors did not suffice.

In this paper, we consider a saliency extension to the standard bag-of-features action detector. One of the reasons that the detection of exchange, bury or haul is hard, is that these actions involve detailed motion patterns and their duration is short. The large part of the total set of features is triggered by irrelevant actions that preceed or follow the detailed action (e.g. walking) or by background clutter (e.g. a person moving in the background). The relevant subset of features is likely to be a small fraction of the total set. To solve this issue, we propose a spatio-temporal saliency map. Its purpose is to improve the selectivity of the feature representation by weighting each feature by its relevance for the action of interest. The spatio-temporal saliency map that we propose, depends on the current video as well as the current action of interest. We provide a simple weighting scheme that is easy to implement, computationally efficient, and deployable for the retrieval/detection of a wide range of actions. We demonstrate that in a bag-of-words setup, the retrieval accuracy can be drastically improved by the proposed spatio-temporal saliency map.

This paper is organized as follows. In Section 2, we summarize related work and indicate which elements we re-use from the action detection literature and where we extend beyond current literature with respect to visual saliency. Section 3 describes our spatio-temporal saliency map and we describe how it can be implemented by the reader in a few simple steps. Section 4 contains the experimental results and we highlight the key improvements with respect to the standard bag-of-features detector. In Section 5 we conclude and summarize our main findings.

## 2. Related Work

### 2.1. Bag-of-Features Action Detectors

The contribution of this paper is a spatio-temporal saliency map for bag-of-features action detectors. Therefore, we summarize the key elements of this class of action detectors here. The bag-of-features model [11] defines a pipeline from features to histograms which are assigned a class label by some classifier. For action detection, the classifier serves simply as a detector that discriminates between the target action vs. the background of negatives. To capture the motion patterns of human actions, STIP features [12] proved to be very effective. For quantization,

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.45

we use a random forest [13]. The final step is the classifier which serves as the action detector. We select the SVM for this purpose, due to its robustness to large feature representations and sparse labels, and the intersection kernel due to its efficiency.

### 2.2. Visual Saliency

Explicit maps of computational saliency have been used to search for the relevant parts in the image to perform visual classification [14,15]. A disadvantage of both methods is that the learned map for a particular class yields the same saliency values for each image. Saliency maps that are both class and image specific were proposed in [16]. Here, the saliency map serves as a weighting function to determine for each image specifically the contribution of features. The rationale in [17] is that both spatial and visual saliency are important and we share that view. Our approach differs in two ways, however. The first difference is that we do not couple the learning of saliency and the learning of the class, because it makes the learning phase and the implementation of the saliency algorithm more complex. We will propose a simple way to learn and compute the saliency. The second difference is that we extend the spatial saliency map to the spatio-temporal domain.

Spatio-temporal regions-of-interest have been addressed in [17]. A foreground region in the space-time volume was identified for each action by applying PageRank on local features in the videos. The rationale was that the background can vary significantly even for the same type of actions in unconstrained videos and that it should be removed. Likewise, in [18], STIP features in the background were removed by an advanced scheme. First in each frame spatial interest points were detected, then background suppression was performed by a center-surround suppression mask. As a final step, local and temporal constraints were imposed to remove the STIP features in the background.

In this paper, we take a different standpoint. We give high weights to relevant features both in the foreground *and* in the background. Hereby we explicitly exploit the property that some features aid in deciding that this video contains the action of interest, and that other features aid in deciding that it does *not* contain the action of interest. We directly exploit the distinctiveness of each of the features in the video, without making assumptions on foreground and background.

## 3. Spatio-Temporal Saliency Map

### 3.1. Individual Feature Relevance for the Action

Our notion of a feature's relevance for the action of interest, is based on three observations. In the explanation and derivation of the equations, we will introduce variables which will keep their meaning throughout this section.

The first observation is our starting point: an SVM classifier can be trained to distinguish an action based on histograms of features from videos. In this paper, the training of a SVM is based on histograms obtained from quantization of STIP features via a random forest. The distances between these histograms are defined by the histogram intersection kernel (see Section 2.1 for motivation):

$$k(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{m} \min\,(p_i, q_i) \qquad (1)$$

with $k$ the kernel, $\boldsymbol{p}$ and $\boldsymbol{q}$ histograms, $i$ the histogram bin, $m$ the number of bins, and $p_i$ and $q_i$ the contents of the $i^{\text{th}}$ histogram bin. Our SVM is clearly a model with an additive kernel.

The second observation is that this SVM model can be rewritten as a combination of support vectors, their coefficients, the histogram of features, and the kernel [19,20]. The decision function of the SVM is:

$$f(\boldsymbol{p}) = b + \sum_{j=1}^{n} \alpha_j \cdot k\,(\boldsymbol{p}, \boldsymbol{z_j}) \qquad (2)$$

where $f$ is the SVM's decision function; with $\boldsymbol{p}$ is the current histogram to be classified, $b$ the bias of the decisio function, $n$ the number of support vectors, $\alpha_j$ the coefficient of the $j^{\text{th}}$ support vector, $k$ the kernel function and $\boldsymbol{z_j}$ is support vector $j$. In the case of an intersection kernel, Equation 2 can be rewritten as [19]:

$$f(\boldsymbol{p}) = b + \sum_{j=1}^{n} \alpha_j \cdot \sum_{i=1}^{m} \min\,(p_i, z_{ji}) \qquad (3)$$

where $z_{ji}$ is the $i^{\text{th}}$ bin of the $j^{\text{th}}$ support vector. Indeed the SVM's decision function is a combination of a bias term, support vectors, their coefficients and the current histogram.

The third observation is that the support for the SVM's decision can be attributed to the bins of the histogram, because the other elements of the linear combination in Equation 2 (i.e. support vectors, coefficients, and kernel) are known after the learning phase [21]. In [21], the support of the histogram bins was defined in a relative manner for visualization purposes only. There, the bias $b$ of the SVM's decision function was not accounted for. In this paper, we need an absolute value for the support of each histogram bin, because we will combine them into a single saliency map. For that purpose, we need to account

for the bias, in such a way that we achieve the following. Our notion of the support is that it is larger than zero if the contribution of the feature is positive (i.e. the action is present) and smaller than zero if negative (absent). Larger support values, positive and negative, indicate more contribution to the respective decision. The support $s(x_i)$ of each histogram bin $i$ for the SVM's decision is defined by:

$$s(p_i) = c(p_i) + \sum_{j=1}^{n} \alpha_j \cdot \min(p_i, z_{ji}) \quad (4)$$

with $z_{ji}$ the value of the $i^{th}$ bin of the $j^{th}$ support vector, and $s(p_i)$ a part of the bias $b$ from Equation 2, under the constraint:

$$\sum_{i=1}^{m} c(p_i) = b \quad (5)$$

Note that indeed the sum over all bins $i$ of Equation 4 equals the SVM's decision function of Equation 3.

There are three regimes to divide the bias $b$ into the $s(x_i)$:

1. uniform across the bins:

$$c(p_i) = b/m \quad (6)$$

2. by prevalence of each bin across the train set:

$$c(p_i) = b \cdot \mu(p_i) \quad (7)$$

with $\mu(p_i)$ the average of the $i^{th}$ bin's values. Because the histograms are normalized to one, the $\mu(p_i)$ across all bins sum to one by construction.

3. by the contents of the histogram bin:

$$c(p_i) = b \cdot p_i \quad (8)$$

where again the $p_i$ sum to one by construction.

Our spatio-temporal saliency map is constructed by assigning to each individual feature in the current video its contribution to the SVM's decision. We call this the support of the individual feature. We can do this because we know which individual feature was quantized into which bin. The support $v$ of a feature $f_{il}$, i.e. the $l^{th}$ feature that was quantized into bin $i$ is:

$$v(f_{il}) = \frac{s(p_i)}{r} \quad (9)$$

with $r$ the total number of features that have been quantized into bin $i$.

With the proposed computation of an individual feature's support, we take advantage of two properties. The first is that we incorporate the SVM's discrimination function, where we take directly advantage of the separating hyperplane between the action and non-action, rather than a derived measure of saliency. Secondly, we extend beyond a single-feature measure of saliency. We take advantage of the full set of features in the current video and the dataset, because we determine the support for the histogram bin first, before assigning the support of individual features. rather than a single-feature based saliency.

## 3.2. Computation of the Saliency Map

The support values $v(f_{il})$ are considered in a scale-space framework [22], to enable reinforcement of salient features that are together in a region. Our expectation is that there may be a foreground (i.e. the action itself) and the background (other activity). If there is such a foreground vs. background distinction, then these become regions-of-interest in the space-time volume. In our saliency map, regions-of-interest are addressed by a gaussian window around the feature points in the space-time volume. This window sums the support of individual features and reinforces salient features that are close together together. The summation is done per feature histogram bin, as we want to avoid regional cancellation that is caused by mixing positive and negative support of different bins. For one individual feature, the summation in the gaussian window is defined by the following equation, where we call the resulting value the weight $w(f_{il})$:

$$w(f_{il}) = v(f_{il}) \cdot \sum_{u=1}^{r} e^{-\left(\left(\frac{x_{il}-x_{iu}}{2\sigma_x}\right)^2+\left(\frac{y_{il}-y_{iu}}{2\sigma_y}\right)^2+\left(\frac{t_{il}-t_{iu}}{2\sigma_t}\right)^2\right)} \quad (10)$$

with the gaussian envelope over the spatial and temporal differences with respect to the point in the space-time volume of the current feature $f_{il}$, where $x$ is the horizontal position, $y$ the vertical position (both in pixels) and $t$ the time (in frames), and $\sigma_x$, $\sigma_y$ and $\sigma_t$ the respective scales. This window accumulates envelope values over all $r$ features (indexed by $u$) that have been quantized into bin $i$.

The spatio-temporal saliency map is a weighting scheme that boosts the relevant features in the feature histogram. As a first step, the weights of the individual features are combined into a single weight $w(p_i)$ for each bin $i$ for each histogram $\boldsymbol{p}$:

$$w(p_i) = \sum_{u=1}^{r} w(f_{iu}) \quad (11)$$

In the final step, each histogram bin $p_i$ is multiplied by its positive weight *w,* by taking the absolute value:

$$p_i' = p_i \cdot |\, w(p_i)| \qquad (12)$$

where $p_i'$ is the new histogram value at bin *i*.

The salient features will boost particular bins in the histogram. For the computation of this spatio-temporal saliency map, there are two parameters: the regime for dividing the bias across the features (Equations 6-8), and the spatio-temporal extent of the map by the size of the gaussian window (Equation 10). With the boosted histograms, a second SVM is trained. For a new test video, a histogram is created by the random forest, then it is boosted using the saliency map, and fed to the second SVM for final classification.

# 4. Evaluation: 48 Actions in 3,480 Movies

## 4.1. Experimental Setup

As a large video database of many diverse and complex human actions, we consider the visint.org database [23]. It contains 3,480 movies of 48 human actions in highly varying settings. The variations are: scenes, recording conditions, viewpoints, persons, and clothing. Each video has been annotated for all 48 actions, where the annotator indicated presence or absence of the action. On average, 7 actions have been indicated to be present in a video. We perform experiments for retrieval of each of the 48 actions.

For each action, we repeat the experiment 5 times, where each repetition uses a randomized train set (50%) and test set (50%). We report the performance on the test set, where we indicate the average and the standard deviation of our performance measure. Our performance measure is Matthews Correlation Coefficient (MCC), because it is independent of the prevalence of an action. The prevalence of the actions varies highly: 'move' occurs in 75.4% of the movies, where 'bury' occurs only in 1.8% of the movies, see column 2 in Table 1. The meaning of the MCC is as follows: a score of 1 (-1) indicates perfect positive (negative) correlation between the action detector and the annotations, where a score of 0 indicates no correlation with the annotations.

The retrieval performance of the standard bag-of-features action detectors is compared against the extended detectors where the saliency map has been added. For both methods, we consider the exact same randomization for each of the 5 repetitions of the retrieval experiment. The parameters of our spatio-temporal saliency map are the bias regime (uniform, prevalence, histogram; see Equations 6-8) and the spatial and temporal scale of the map (Equation 10). The spatial scales are isotropic: $\sigma_{xy}=\sigma_x=\sigma_{y,}$ because different scales in x- and y-direction not significantly impact the results. The scales are varied: $\sigma_{xy}$ = [1, 5, 10] pixels, and $\sigma_t$ = [1, 5, 10] frames.

## 4.2. Organization of the Results

The retrieval results for the human actions from visint.org are summarized in Table 1. The first column lists the actions. The prevalence of the action in the database is indicated in the second column; it is clear that the prevalence of actions varies severely, from as few as 1.8% ('bury'). The third column indicates the median number of STIP features found in the movies where the action occurs; clearly some actions contain few features ('give') due to their short duration and subtle motion, where others trigger many features ('haul'). The performance of the standard bag-of-features action detectors is listed in column 4. The performance of the extended detectors tat include the spatio-temporal saliency map, is listed in column 5. Columns 6 and 7 indicate the spatial and temporal scale of the saliency map (see Section 3.2), and column 8 indicates how the bias has been spread over the features (see Section 3.1). We have varied the scale and bias parameters and report the best result here. Column 8 contains the merit that is gained by extending the action detectors by the spatio-temporal saliency map.

## 4.3. Findings

The first finding from Table 1 is that many actions could not be retrieved at all by the standard bag-of-features action detectors – yet these can be retrieved with reasonable precision when the saliency map is added. These actions are: attach (0.09, was 0.00), catch (0.11, was 0.01), exchange (0.13, was 0.00), get (0.10, was 0.01), hand (0.14, was 0.01), haul (0.18, was 0.00), hit (0.14, was 0.00), kick (0.21, was 0.03), push (0.09, was 0.02), putdown (0.13, was 0.00), replace (0.17, was 0.03), snatch (0.12, was 0.00). This result is important, as these are exactly the interesting yet complex actions that involve interactions with items in the environment.

The second finding is that 40 out of the 48 actions are improved, and that 25 are improved significantly (MCC increase > 0.05), whereas the degradations for 8 out of 48 actions are not significant (MCC decrease < 0.05). The degradations occur systematically for the actions that already had a good performance without saliency map. We conclude that the saliency maps achieve a systematic improvement across the board and significant improvements for 25 actions.

TABLE 1
MERIT OF SPATIO-TEMPORAL SALIENCY FOR DISCRIMINATION OF HUMAN ACTIONS

| Human Action | Prev. | Feat. | Standard | Saliency | $\sigma_{xy}$ | $\sigma_t$ | Bias | Merit |
|---|---|---|---|---|---|---|---|---|
| Attach | 7.5% | 232 | 0.00±0.02 | **0.09±0.03** | 1 | 1 | uniform | **+0.09** |
| Bounce | 8.8% | 282 | 0.09±0.02 | **0.15±0.09** | 1 | 1 | histogram | **+0.06** |
| Bury | 1.8% | 242 | 0.14±0.12 | **0.22±0.09** | 1 | 1 | prevalence | **+0.08** |
| Catch | 5.0% | 260 | 0.01±0.02 | **0.11±0.07** | 1 | 1 | uniform | **+0.10** |
| Close | 6.4% | 200 | 0.06±0.10 | **0.16±0.03** | 1 | 1 | uniform | **+0.10** |
| Collide | 11.5% | 256 | 0.08±0.09 | **0.17±0.01** | 1 | 1 | uniform | **+0.09** |
| Enter | 16.3% | 392 | 0.06±0.02 | **0.20±0.02** | 1 | 1 | uniform | **+0.14** |
| Exchange | 4.3% | 223 | 0.00±0.02 | **0.13±0.02** | 1 | 1 | uniform | **+0.13** |
| Exit | 12.8% | 461 | 0.17±0.03 | **0.27±0.03** | 1 | 1 | uniform | **+0.10** |
| Flee | 5.2% | 418 | 0.19±0.03 | **0.28±0.02** | 5 | 5 | uniform | **+0.09** |
| Get | 13.0% | 261 | 0.01±0.04 | **0.10±0.03** | 10 | 5 | histogram | **+0.09** |
| Give | 7.4% | 150 | 0.08±0.01 | **0.16±0.04** | 1 | 1 | prevalence | **+0.08** |
| Hand | 7.0% | 139 | 0.01±0.02 | **0.14±0.05** | 1 | 1 | histogram | **+0.13** |
| Haul | 5.1% | 488 | 0.00±0.02 | **0.18±0.05** | 1 | 1 | histogram | **+0.18** |
| Hit | 11.1% | 256 | 0.00±0.03 | **0.14±0.05** | 1 | 10 | histogram | **+0.14** |
| Kick | 3.7% | 261 | 0.03±0.05 | **0.21±0.07** | 1 | 1 | uniform | **+0.18** |
| Push | 14.5% | 306 | 0.02±0.03 | **0.09±0.01** | 1 | 1 | uniform | **+0.07** |
| Putdown | 11.6% | 237 | 0.00±0.03 | **0.13±0.01** | 5 | 1 | uniform | **+0.13** |
| Receive | 11.5% | 164 | 0.08±0.07 | **0.15±0.02** | 1 | 1 | histogram | **+0.07** |
| Replace | 5.0% | 413 | 0.03±0.06 | **0.17±0.05** | 5 | 1 | prevalence | **+0.14** |
| Snatch | 9.2% | 189 | 0.00±0.03 | **0.12±0.03** | 1 | 1 | uniform | **+0.12** |
| Take | 19.9% | 190 | 0.04±0.01 | **0.15±0.02** | 1 | 1 | uniform | **+0.11** |
| Throw | 6.1% | 231 | 0.05±0.04 | **0.11±0.03** | 5 | 5 | uniform | **+0.06** |

*Section 4.2 explains the columns. Section 4.3 highlights the most prominent findings.*

The third finding is that by adding the saliency map 13 actions now achieve a reasonable retrieval performance. To that end, we consider actions that are improved significantly (MCC increase > 0.05) and that now get a score of MCC > 0.20. These actions are: bury, enter, exit, flee, kick. Actions that have improved significantly and now get a score of MCC > 0.15 are: bounce, close, collide, give, haul, receive, replace, take. These actions have in common that they are hard to recognize because they involve subtle motion and have a short duration. Moreover, they typically have low prevalence in the dataset so there are relatively few examples to learn from. Finally, they involve interactions with usually small items which are hard to detect.

The fourth finding is that the spatio-temporal extent of the map, by means of the gaussian windows, matters. For 17 out of 48 actions, the scale in either the spatial and/or temporal dimension is ≥ 5 pixels or frames.

The fifth finding is that the average improvement across the board is 35.3%, from an average MCC = 0.16 for the standard bag-of-features action detectors, to MCC = 0.22 with inclusion of the proposed saliency map.

### 4.4. Qualitative Examples

Equation 9 quantifies the contribution of an individual feature for the current action of interest. An example is shown in Figure 1. Horizontal upward motion is an important cue (red) for 'jump', where it is a negative cue (blue) for 'move'. For examples of saliency of motion features for the IXMAS dataset of 12 human actions, see:
http://www.youtube.com/intelligentImaging

Figure 1: Saliency of STIP features for the same movie of a jumping person, for actions 'jump' (left) and 'move' (right).

## 5. Conclusions

This paper has addressed the recognition of videos that contain a particular human action. Recognizing complex actions, e.g. a person who is replacing an item, is hard due to short duration and subtle motion. Additionally, such complex actions do not occur often, which leads to a small set of positive samples, further complicating the learning and retrieval. These challenges imply a need to find the relevant features in the midst of all features in a video fragment. As a solution, we have proposed a spatio-temporal saliency map to increase the selectivity of the feature representation. In our experiments, we have shown the merit of this map for the retrieval of 48 human actions in 3,480 movies, demonstrating a systematic improvement across the board of 35.3% on average and significant improvements for 25 out of 48 actions.

## References


[1] T. Guha, R.K.Ward, "Learning Sparse Representations for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2012.

[2] C. Schuldt, I. Laptev, B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *International Conference on Pattern Recognition,* 2004.

[3] L. Gorelick, M. Blank, E. Shechtmanm, M. Irani, R. Basri, " Actions as Space-Time Shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2007.

[4] S. Sadanand, J. J. Corso, "Action bank: A high-level representation of activity in video," *International Conference on Computer Vision and Pattern Recognition,* 2012.

[5] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, "Learning Realistic Human Actions from Movies," *International Conference on Computer Vision and Pattern Recognition,* 2008.

[6] G.J. Burghouts, K. Schutte, "Correlations between 48 human actions improve their performance," *International Conference on Pattern Recognition*, 2012.

[7] G.J. Burghouts, K. Schutte, "Spatio-Temporal Layout of Human Actions for Improved Bag-of-Words Action Detection," Pattern Recognition Letters, 2013.

[8] G.J. Burghouts, K. Schutte, H. Bouma, R.J.M. den Hollander, "Selection of Negative Samples and Two-Stage Combination of Multiple Features for Action Detection in Thousands of Videos," Machine Vision and Applications, 2013.

[9] P. Hanckmann, K. Schutte, G.J. Burghouts, "Automated Textual Descriptions for a Wide Range of Video Events with 48 Human Actions," *European Conference on Computer Vision*, 2012.

[10] G.J. Burghouts, "Soft-Assignment Random-Forest with an Application to Discriminative Representation of 48 Human Actions in Videos", International Journal of Pattern Recognition and Artificial Intelligence, 2013.

[11] J. Sivic, A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *International Conference on Computer Vision,* 2003.

[12] I. Laptev, " On Space-Time Interest Points," *International Journal of Computer Vision,* 2005.

[13] F. Moosmann, B. Triggs, F. Jurie, "Randomized Clustering Forests for Building Fast and Discriminative Visual Vocabularies," *Neural Information Processing Systems,* 2006.

[14] T. Harada, Y. Ushiku, Y. Yamashita, Y. Kuniyoshi, "Discriminative spatial pyramid," *International Conference on Computer Vision and Pattern Recognition,* 2011.

[15] B. Yao, A. Khosla, L. Fei-Fei, " Combining randomization and discrimination for fine-grained image categorization," *International Conference on Computer Vision and Pattern Recognition,* 2011.

[16] G. Sharma, F. Jurie, C. Schmid, "Discriminative Spatial Saliency for Image Classification," *International Conference on Computer Vision and Pattern Recognition,* 2012.

[17] J. Liu, J. Luo, M. Shah, "Recognizing Realistic Actions from Videos 'in the Wild'," *International Conference on Computer Vision and Pattern Recognition,* 2009.

[18] B. Chakraborty, M. Holte, T.B. Moeslund, J. Gonzalez, "Selective Spatio-Temporal Interest Points," *Computer Vision and Image Understanding,* 2012.

[19] S. Maji, A.C. Berg, J. Malik, "Classification using intersection kernel support vector machines is efficient," *International Conference on Computer Vision and Pattern Recognition,* 2008.

[20] A. Vedaldi, A. Zisserman, "Efficient additive kernels via explicit feature maps," *International Conference on Computer Vision and Pattern Recognition,* 2010.

[21] J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha, "The Visual Extent of an Object – Suppose We Know the Object Locations," *International Conference on Computer Vision,* 2012.

[22] J.J. Koenderink, "The structure of images," *Biological Cybernetics*, 1984.

[23] www.visint.org, development kit for recognition task.