Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 83–89

ISBN 978-83-60810-22-4

Automatic Image Annotation by Image Fragment Matching

Mariusz Paradowski
Institute of Informatics
Wroclaw University of Technology, Poland
Email: mariusz.paradowski@pwr.wroc.pl

Andrzej Śluzek
Khalifa University, Abu Dhabi
United Arab Emirates

***Abstract*—Automatic image annotation problem is one of the most important sub-problems of computer vision. It is strongly related to and goes beyond image recognition. The key goal of annotation is to assign a set of words from a given dictionary to a previously unseen image. In this paper, we address two key problems related to image annotation. The first one is low precision of generated answers, the second one is automatic localization of image fragments related to annotations. The proposed method utilizes image fragment matching to precisely localize near-duplicate visual content of images.**

I. Introduction

AUTOMATIC image annotation (*AIA*) is an extension of image recognition. The input for an AIA method is an image. The output is a set of words (also referred as classes), from a given dictionary, which describe the input image in a best possible way. However, in the AIA problem additional assumptions are made [1], comparing to the classic recognition problem [2]. The major assumption is the lack of relations between the words and their visual representation (feature vectors) within a single image. Each word within the image description is represented by all feature vectors, which can be generated from the visual content of the image. Instead of a one to one relations (like in a classic recognition), there are many to many relations.

Most of the available image annotation methods outputs a set of words, without any information on the localization of the detected words. Precise localization of image fragments related to detected words is a challenging task. The earliest approaches to AIA assumed image segmentation. Each segment has been related with one or more words, e.g. [3], [4]. Such approach fails in most cases, because image segmentation usually do not covers true boundaries of objects represented by the dictionary. More recent research shows that image segmentation used for the purpose of annotation do not have to recreate true object boundaries, e.g. [5]. Rectangular blocks [5], [8] or various setups of fixed regions [6] became very popular and effective. Obtained quality results are usually better than those with classic segmentation, but word localization is practically impossible. There are attempts to handle AIA with word localization without image segmentation, e.g. [7]. A very dense grid of rectangles is utilized. In general, all mentioned approaches give some rough, very imprecise information on the localization of word representations on the image.

Another problem is a low precision of practically all available image annotation approaches. Even the state-of-the-art methods (e.g. [8]) tested on relatively simple image databases fail to reach precision above 90%. There are many sources of this problem, however one of the most prominent is the mentioned lack of relations between words and image fragments. Construction of high quality recognition models is not possible, due to very high noise in the training data. Only a small fragment of positive training data is correct [7].

Recent research in AIA (e.g. [8]) shows that distance based methods, strongly related to image retrieval, give very promising results. Key-point based, near-duplicate retrieval is particularly interesting. Key-points such as *SIFT* [9] are used to model local photometric properties of images. These fragments are matched together to establish candidate relations between images. Global geometric [10], [11] or topological [12] constraints provide necessary information to verify the correctness of candidates.

The paper presents a new automatic image annotation method, strongly related to key-point based image retrieval. It solves mentioned problems for a specific sub-domain of images and words. Generated annotations reach high precision. The method it is able to precisely locate words on the image, without such information in the training set. The proposed approach has limited applicability, it is only able to annotate words with near-duplicate visual representation.

II. Proposed approach

The proposed method is called *Automatic Image Annotation by Matching* (AIAM). Let us formalize important concepts at the beginning. We follow with a general description and all the necessary details.

A. Formal definitions

Let us give the formal definitions of important concepts. First, we focus on automatic image annotation. Later on, near-duplicate fragment detection is described.

a) Automatic image annotation: The dictionary $\mathcal{W} = \{w_1, w_2, ..., w_k\}$ is a set of words $w_x : x = \{1, ..k\}$, where k is the size of the dictionary. Words are strings, they do not have any related semantics. The training set $\mathcal{I}$ consists of n pairs: images I_x and their annotations $W_x \subseteq \mathcal{W}$, where $x = \{1, ..., n\}$:

$$\mathcal{I} = \{(I_1, W_1), (I_2, W_2), ..., (I_n, W_n)\}. \quad (1)$$

Input image J is an image to be described in an automatic manner. Automatic image annotation method $\mathcal{A}$ describes image J using words from the dictionary $\mathcal{W}$, based on the data from the training set $\mathcal{I}$. Output of the method is an annotation $\mathcal{A}(J) = W_J \subseteq \mathcal{W}$ of the input image J.

b) Near-duplicate image fragments: Near-duplicate image fragment matching method $\mathcal{M}$ accepts a pair of images (X, Y) as an input. If these two images contain near-duplicate image fragments, they are marked using outlines (convex hulls). Each pair of outlines represent a single near-duplicate image fragment. The output is a set of m outline pairs. Both the number of outline pairs (m) and the outlines are detected fully automatically by the matching method $\mathcal{M}(X, Y)$:

$$\mathcal{M}(X, Y) = \left\{(f_X^1, f_Y^1), (f_X^2, f_Y^2), ..., (f_X^m, f_Y^m)\right\}. \quad (2)$$

B. Outline of the method

The proposed automatic image annotator utilizes near-duplicate fragment matching method as a key component. Matching of near-duplicates has to be performed with possible highest quality. Out of several available approaches to image fragment matching, a method proposed by the authors is chosen [13]. The method utilizes low level image features (*SIFT* [9]) and affine geometry. A six dimensional probability density function of available affine transformations is constructed. The density function is modelled using a non-parametric approach (sparse histogram built using hash-table), which allows simultaneous matching of unknown number of image fragments.

The largest advantage of this matching method is high precision. There are only a few *false positive* errors. *False negatives* are much frequent ones, but lack of detection is a problem of much less grade. Additionally, the applied method generates highly precise outlines of detected near-duplicates.

In the first step of the method input image J is matched with all images $I_x : x = \{1, ..., n\}$ from the training set $\mathcal{I}$. Output of a single matching is a set of outlines representing near-duplicates. Exemplary near-duplicates for a selected input image J are shown in Fig. 1.

Generated set of near-duplicates may not be directly used to propagate words from image annotations. There are no relations between image fragments and words. Presence of near-duplicates is only a prerequisite to word propagation. First, it has to be determined which words are to be propagated. Second, matching may be erroneous, thus it has to be verified.

1) Propagation of words: For each image I_x from the training set $\mathcal{I}$ we have to automatically determine additional information required for word propagation. A set of images $S_x^w \subset \mathcal{I}$ is called a *supporting set* for word $w \in W_x$ and image I_x. Supporting set S_x^w contains images having exactly one common word (w) with annotation W_x of image I_x. Formally, supporting set S_x^w for image I_x and word $w \in W_x$ is defined as:

Fig. 1. Examples of matched near-duplicates for input image J and various images from the training set $\mathcal{I}$.

$$S_x^w = \bigcup_{\substack{1 \le y \le n \\ y \ne x}} \{I_y\} : (W_y \cap W_x = \{w\}). \quad (3)$$

Supporting set S_x^w for a given image I_x and word $w \in W_x$ allows to:

- propagate word w from the annotation W_x with higher precision,
- relate word w with matched image fragments generated by $\mathcal{M}(J, I_x)$,
- verify the correctness of image fragments matching $\mathcal{M}(J, I_x)$.

In case the fragment matching of two images (J and I_x) is not empty ($\mathcal{M}(J, I_x) \neq \emptyset$), propagation is performed for all words that annotate image I_x. Input image J is matched with all images from the supporting sets ($\mathcal{M}(J, s) : s \in S_x^w$). If a word $w \in W_x$ has supporting matches ($\mathcal{M}(J, s) \neq \emptyset$), it is verified if matching outlines intersect each other. To get the most credible answer (we assume that matching precision is high, but recall may be low), the largest intersection is selected:

$$(f_J^1, f_{I_x}^1), (g_J^1, g_s^1) = \underset{\substack{(f_J, f_{I_x}) \in \mathcal{M}(J, I_x) \\ (g_J, g_s) \in \mathcal{M}(J,s)}}{\operatorname{argmax}} |f_J \cap g_J|. \quad (4)$$

Near-duplicate fragments generated by method $\mathcal{M}$ are considered as valid, if and only if the intersection of outlines is large enough, i.e. it meets the following criterion (*first intersection test*):

$$\left(\frac{|f_J^1 \cap g_J^1|}{|f_J^1|} > t \right) \cap \left(\frac{|f_J^1 \cap g_J^1|}{|g_J^1|} > t \right), \quad (5)$$

where: $t \in \langle 0, 1)$ is the method parameter. This concept is presented in Fig. 2.

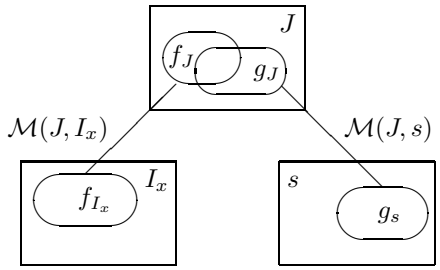


Fig. 2. Basic propagation of words with the usage of supporting sets. There are two different near-duplicate sets on image J, the first one comes from $\mathcal{M}(J, I_x)$, the second one from $\mathcal{M}(J, s)$. Outline intersections on image J have to match as close as possible.

Basic propagation of words with the proposed intersection tests eliminates a large number of *false positives*. However, some of them are still accepted as correct matches.

2) Extended propagation of words: Let us now propose another variant of word propagation. To get even higher precision of annotations, we introduce an extended propagation method. It is also based on the image fragment matching and the sets of supporting images S_x^w. It encapsulates the *first intersection test* used in basic propagation (eq. 5).

The extended propagation also utilizes a triple of images: J, I_x and a supporting image s. As already mentioned, the basic propagation may not be sufficient to prevent *false positives*. Given the results of matching $\mathcal{M}(I_x, s)$, it is possible to define two additional tests of propagation correctness. They allow for even better rejection of incorrect matches. In case $\mathcal{M}(I_x, s) \neq \emptyset$, largest intersections of outlines within images I_x and s are found. The first one is done on image I_x:

$$(f_J^2, f_{I_x}^2), (h_{I_x}^2, h_s^2) = \underset{\substack{(f_J, f_{I_x}) \in \mathcal{M}(J, I_x) \\ (h_{I_x}, h_s) \in \mathcal{M}(I_x, s)}}{\operatorname{argmax}} |f_{I_x} \cap h_{I_x}|, \quad (6)$$

The second one is done on image s (from the supporting set S_x^w):

$$(g_J^3, g_s^3), (h_{I_x}^3, h_s^3) = \underset{\substack{(g_J, g_s) \in \mathcal{M}(J,s) \\ (h_{I_x}, h_s) \in \mathcal{M}(I_x, s)}}{\operatorname{argmax}} |g_s \cap h_s|. \quad (7)$$

Having the largest intersections of outlines, we propose two additional tests to validate the correctness of $\mathcal{M}(J, I_x)$ matching. For each test a threshold $t \in \langle 0, 1)$ is given (the same as in basic propagation, see eq. 5). The first test examines the intersection of outlines in image I_x:

$$\left(\frac{|f_{I_x}^2 \cap h_{I_x}^2|}{|f_{I_x}^2|} > t \right) \cap \left(\frac{|f_{I_x}^2 \cap h_{I_x}^2|}{|h_{I_x}^2|} > t \right), \quad (8)$$

The second test examines the intersection of outlines in image s:

$$\left(\frac{|g_s^3 \cap h_s^3|}{|g_s^3|} > t \right) \cap \left(\frac{|g_s^3 \cap h_s^3|}{|h_s^3|} > t \right). \quad (9)$$

The idea is presented in Fig. 3.

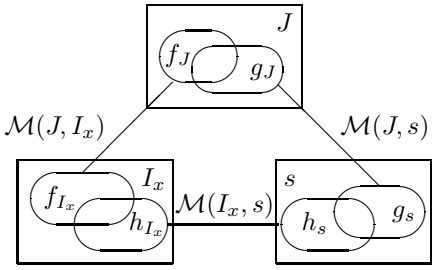


Fig. 3. Extended word propagation routine. Three outline intersection tests are performed. The first one is done on image J (matches $\mathcal{M}(J, I_x)$ and $\mathcal{M}(J, s)$), the second one is done on image I_x (matches $\mathcal{M}(J, I_x)$ and $\mathcal{M}(I_x, s)$) and the last one on image s (matches $\mathcal{M}(J, s)$ and $\mathcal{M}(I_x, s)$).

Experimental results presented in the later part of the paper shown that the extended propagation routine is fully sufficient. Less precise matching routines may require additional tests to reject false matches. In case such tests are necessary, they may be constructed using the same idea, as the above.

C. Localization of annotated objects

One of our goals is to precisely localize fragments of images to which generated annotation should be related. It is possible due to the application of image fragment matching routine. Presented word propagation mechanism automatically relates words with some image fragments.

The input image J is automatically annotated with word $w \in \mathcal{W}$. The word could be propagated from any image $I_x \in \mathcal{I}$ annotated by this word ($w \in W_x$). To maximize recall we should propagate outlines from all matched images. However, such solution is not a precise and valid one.

Let us assume that three (or more) images from the training set contain a visual representation of word w on a near-duplicate background. Image fragment matching method $\mathcal{M}$ works on a purely visual manner, and thus does know nothing about visual representation of word w or any neighboring background. The main near-duplicate match $\mathcal{M}(J, I_x)$ and the supporting matches $\mathcal{M}(J, s)$ and $\mathcal{M}(I_x, s)$ are going to generate outlines containing both the proper image fragment

representing word w and the neighboring background. Intersection tests (see eqs. 5, 8 and 9) are also going to accept both the object and the background.

However, the intersection tests accept only near-duplicates of a similar relative fragment size (up to a threshold). If $\mathcal{M}(J, I_x)$ returns both the objects of interest and the background, only a small subset (images containing both the object and the background) of supporting set will support this match. Thus it is worth rejecting near-duplicate matches $\mathcal{M}(J, I_x)$ supported by small subsets of their supporting sets. We propose to select a single image I_x, for which the matched subset of the supporting set is the largest. In other words, to get the most credible match I_x, it is expected that the match is supported by as many other matches as possible.

D. The dictionary and the training set

The proposed automatic image annotation approach does have three noticeable flaws. They are related to the training set and the dictionary. We are going to describe them in detail.

First, the method annotates words that have near-duplicate visual representation. The proposed annotator utilizes highly precise, key-point based, image fragment matching method. The matching method is only able to detect near-duplicates, all other types of visual similarity are not taken into account. Words such as *sky*, *cloud*, etc. are not going to be detected (recall will be near zero or zero). The proposed method is applicable to words representing mainly *rigid bodies*, similar in shape and appearance. Suprizingly, most of automatic image annotation methods usually fail to correctly annotate such words. They focus on the mentioned holistic concepts, which can be easily generalized. This method does the opposite, it does not generalize. It annotates words with complex, but highly repeatable appearance.

Second, effective annotation of a word requires a sufficient representation in the training set. Propagation of a word requires an existence of the supporting set (see II-B1). The larger the supporting set is, the easier it is to propagate annotations. Used near-duplicate detection method has very high precision, but lower recall. It rarely makes *false positive* errors, but often causes *false negatives*. False negatives may be eliminated by providing a larger set of examples, e.g. captured under various lighting conditions, relative position to camera, etc. To propagate a word, it is sufficient to have only three correct matches. Larger word representations make finding such triples much easier.

Third, background repetition in the training set should be minimized. Near-duplicate detection is purely visual, i.e. it does not differentiate between objects of interest and background. If the same background goes together with object of interest in most images, it is also considered a valid part of image. Generated annotation are going to be correct. However, localization of these annotations on the image will not depict the true and expected boundaries. We consider this flaw the most prominent one, because it may require manual modification of the training set.

III. Experimental verification

The goal of experimental verification is to check the expected properties of the proposed method. According to the initial assumptions, the main expectation is a high precision of generated annotations. Additionally, annotated objects have to be precisely localized.

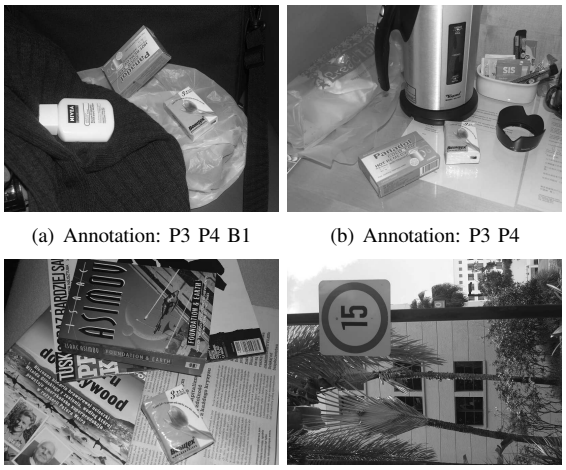
Presented experimental verification consists of four parts:

- presentation of exemplary results of the method,
- setup of method parameters,
- analysis of annotation quality,
- analysis of object localization quality.

The image set consists of 100 images, together with annotations and outlines for each annotated word[1]. Single images contain multiple objects of interest, thus the set is well suited for AIA. The dictionary is constructed from all available annotations. All performed tests are done using *leave one out* experimental protocol. Typical, state-of-the-art automatic image annotation methods fail to get high precision results.

A. Exemplary results

According to the definition of automatic image annotation, we have a database of annotated images. Annotations are related with entire images, there are no relations between words and image fragments. Exemplary elements from the training set are presented in Fig. 4.



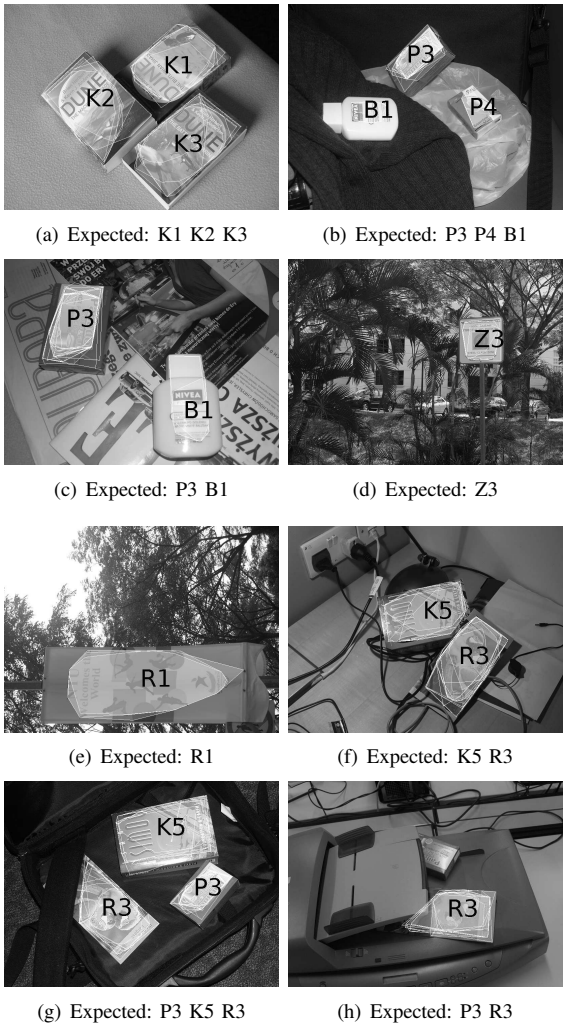
(a) Annotation: P3 P4 B1 (b) Annotation: P3 P4

(c) Annotation: P4 P5 (d) Annotation: Z1

Fig. 4. Exemplary described images from the training set. According to the definition of automatic image annotation problem, annotations are related with entire images, instead of image fragments.

Fig. 5 presents exemplary results generated by the proposed method. The method generates image annotations and outlines of image fragments representing single words. Annotation outlines are generated by the near-duplicate image fragment detection method. They are a combination of at least two (usually much more) near-duplicate outlines. Of course, the more precise the near-duplicate outlines are, the better are the final results.

[1]http://www.ii.pwr.wroc.pl/∼visible



(a) Expected: K1 K2 K3 (b) Expected: P3 P4 B1

(c) Expected: P3 B1 (d) Expected: Z3

(e) Expected: R1 (f) Expected: K5 R3

(g) Expected: P3 K5 R3 (h) Expected: P3 R3

Fig. 5. Generated annotations together with localization of found words. Expected words are shown below the images. It should be noted, that the training set does not contain any information on words localization. It is determined by the proposed method fully automatically.

The automatic image annotator uses geometric approach to image fragment matching. The geometric approach tends to generate outlines smaller than the true boundaries of objects. Due to usage of key-points (*SIFT*) as the elementary visual data, it is very difficult to recreate affine transformations on object boundaries. Simply, there are not enough key-points on the boundaries (usually there are very few). Key-points outside the true boundaries may not be taken into consideration, because they do not generate valid affine transformations. Mentioned behaviour may be observed in Fig. 5.

B. Setup of method parameters

The second part of experimental verification addresses the parameter setup. Two standard annotation quality measures are used: *recall* and *precision*. Unlike in automatic image annotation, these quality measures may be calculated in two different manners. The first one is the classic approach, based on the presence or absence of words in annotations. The second one is based on the quality of generated outlines. Outline pixels on the image may by considered as false positives, false negatives and true positives (true negatives are irrelevant) and used in complex measures: *recall* and *precision*.

TABLE I
QUALITY OF RESULTS FOR VARIOUS SETUP OF METHOD PARAMETERS. EXTENDED PROPAGATION PROVIDES HIGHER PRECISION COMPARING TO BASIC PROPAGATION.

thres. t	*prec.* [obj.]	*recall* [obj.]	*prec.* [area]	*recall* [area]
Basic propagation				
0.00	0.92	0.81	0.86	0.43
0.25	0.95	0.80	0.89	0.43
0.33	0.96	0.80	0.89	0.43
0.50	0.96	0.80	0.90	0.43
0.66	0.96	0.80	0.89	0.43
0.75	0.96	0.78	0.89	0.43
Extended propagation				
0.00	0.99	0.69	0.95	0.40
0.25	1.00	0.68	0.96	0.38
0.33	1.00	0.67	0.95	0.39
0.50	1.00	0.66	0.96	0.40
0.66	1.00	0.62	0.96	0.37
0.75	1.00	0.58	0.97	0.36

Two variants of the proposed method are taken into consideration: basic propagation (see Sec. II-B1) and extended propagation (see Sec. II-B2). Results achieved for the basic propagation reach precision near 95% and recall near 80%. However, the key goal is to get as high precision as possible. Given the extended propagation, precision is equal to 100%. Recall is lowered to 68% because higher requirements of the propagation routine. Summary of the method quality is presented in Tab. I.

Cutoff threshold t is the second parameter of the method. The threshold is used in the supporting set acceptance tests (eqs. 5, 8 and 9). The larger the threshold, the more similar size of the detected fragments is required. According to the performed tests, $t = 0.25$ is sufficient. However, given the need of background matches rejection (see Sec. II-D), suggested threshold value is higher and equal to $t = 0.50$.

C. Quality of generated annotations

The third part of the experimental verification is the analysis of annotation quality. Table II contains detailed results calculated for each word from the dictionary.

Taking into account the initial requirements, the most interesting column is the *false positives* one. No *false positives* are found during our experiments with the extended propagation. Proposed extended propagation routine with three built-in validation tests is sufficient to eliminate *false positive* matches. The quality of generated outlines plays the key role. Better outlines cause larger outline intersections with smaller relative size changes. Better intersections are a better confirmation of near-duplicate matches, and a better confirmation of generated annotations.

The second interesting observation is related to *false negatives*. Words with the lowest recall (mostly 0) are the least

TABLE II
QUALITY OF GENERATED ANNOTATIONS. ONLY THE PRESENCE OF WORDS IS TAKEN INTO ACCOUNT. LOCALIZATION OF WORDS IS NOT TAKEN INTO ACCOUNT.

class	*TP*	*FP*	*FN*	*prec.*	*recall*
B1	7	0	0	1.00	1.00
B2	0	0	5	–	0.00
B3	0	0	3	–	0.00
H1	3	0	1	1.00	0.75
H2	0	0	2	–	0.00
HD	0	0	2	–	0.00
IP	0	0	2	–	0.00
K1	12	0	0	1.00	1.00
K2	3	0	4	1.00	0.42
K3	7	0	0	1.00	1.00
K4	0	0	3	–	0.00
K5	7	0	0	1.00	1.00
P1	7	0	1	1.00	0.87
P2	0	0	2	–	0.00
P3	20	0	1	1.00	0.95
P4	2	0	5	1.00	0.28
P5	5	0	1	1.00	0.83
R1	4	0	0	1.00	1.00
R2	0	0	2	–	0.00
R3	15	0	0	1.00	1.00
T1	0	0	3	–	0.00
T2	0	0	3	–	0.00
Z1	2	0	3	1.00	0.40
Z2	4	0	0	1.00	1.00
Z3	5	0	0	1.00	1.00
Z4	0	0	4	–	0.00
Z5	0	0	2	–	0.00
Z6	0	0	2	–	0.00
Z7	0	0	2	–	0.00
all	103	0	53	1.00	0.66

frequent ones in the training set. This confirms the second mentioned flaw of the proposed annotator (see Sec. II-D). In case of more frequent words, recall grows up. A sufficiently large number of diversified training examples makes near-duplicate matching possible.

D. Quality of annotated words localization

The last part of the presented experimental verification is assessment of object localization quality. The assessment requires additional information, *not available* in the training set, i.e. outlines of all annotated objects. This allows to verify the quality of localization up to single pixels.

The quality measurement is performed using *precision* and *recall* based on outline intersections. Detailed results are shown in Tab. III. Precision of results is slightly lower than 100%. It is directly related to the image fragment matching method: outlines of generated fragments are in sometimes larger than the true boundaries of objects. Annotation outline is a sum of several matching outlines, i.e. it contains all false positives of these outlines. One of possible solutions is to use intersection of all outlines, however this causes a very large drop of recall.

IV. SUMMARY

A new method of automatic image annotation is presented. The method is called *Automatic Image Annotation by Matching* (AIAM). To annotate a previously unseen image, word

TABLE III
QUALITY OF GENERATED OUTLINES. QUALITY MEASURES ARE CALCULATED USING EXPECTED AND GENERATED OUTLINE INTERSECTIONS.

class	*prec.*	*recall*	*class*	*prec.*	*recall*
B1	0.91	0.35	B2	–	0.00
B3	–	0.00	H1	0.99	0.64
H2	–	0.00	HD	–	0.00
IP	–	0.00	K1	0.99	0.86
K2	0.98	0.30	K3	0.91	0.85
K4	–	0.00	K5	0.94	0.91
P1	0.90	0.48	P2	–	0.00
P3	0.97	0.78	P4	0.99	0.09
P5	0.98	0.68	R1	0.98	0.63
R2	–	0.00	R3	0.99	0.85
T1	–	0.00	T2	–	0.00
Z1	1.00	0.04	Z2	1.00	0.17
Z3	0.99	0.74	Z4	–	0.00
Z5	–	0.00	Z6	–	0.00
Z7	–	0.00			
all				0.96	0.49

propagation is used. Words are propagated between images sharing near-duplicate visual content. Near-duplicates are detected using low level key-points (*SIFT*) and global affine geometry.

Two routines of near-duplicate image matching verification are proposed: basic and extended one. Triples of images sharing common visual content are found. Each accepted triple is a sufficient premise to propagate a single word from the training set image into the final annotation.

For a sub-domain of images (with the presence of near-duplicates) the method is able to get very high precision (up to 100%). Recall achieved for the test image set reaches 68%. Another interesting feature of the proposed annotator, is the ability to localize fragments of image representing annotated words. The major feature and the limitation of the proposed approach is the limitation to near-duplicates. Words with very similar visual appearance may be used in the dictionary. All other words will be automatically discarded, because near-duplicate matches will not be found.

Further research will cover two main areas. The first one is the speed-up of matching routines, while keeping high quality of outlines (partially done already). The second one will focus on increase of recall, within the near-duplicate detection framework. Once there two goal are reached, we expect to get a much more viable annotator.

REFERENCES

[1] P. Duygulu, K. Barnard, N. de Freitas and David Forsyth, Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, Proceedings of Seventh European Conference on Computer Vision (ECCV'02), vol. 4, pp. 97-112, 2002.

[2] M. Kurzynski, Objects Recognition: statistical methods (in Polish), Wroclaw University of Technology Publishers, 1997.

[3] J. Jeon, V. Lavrenko, R. Manmatha, Automatic Image Annotation and Retrieval using Cross-Media Relevance Models, Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 119-126, 2003.

[4] V. Lavrenko, R. Manmatha, J. Jeon, A Model for Learning the Semantics of Pictures, Proceedings of NIPS, MIT Press, 2003.

[5] S. L. Feng, R. Manmatha, V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), Vol. 2, pp. 1002-1009, 2004.

[6] B. Shah, R. Benton, Z.H. Wu and V. Raghavan, Automatic and Semi-Automatic Techniques for Image Annotation, Semantic Based Visual Information Retrieval, pp. 112-134, 2007.

[7] G. Carneiro and N. Vasconcelos, Formulating Semantic Image Annotation as a Supervised Learning Problem, Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, pp. 163-168, 2005.

[8] M. Stanek, B. Broda and H. Kwaśnicka, PATSI – Photo Annotation through Finding Similar Images with Multivariate Gaussian Models, Proceedings of the ICCVG (2), pp. 284-291, 2010.

[9] D. G. Lowe, Object recognition from local scale-invariant features, Proc. 7th IEEE Int. Conf. Computer Vision, Vol. 2, pp. 1150–1157, 1999.

[10] H. Jégou, M. Douze and C. Schmid, Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search, Proceedings of the 10th European Conference on Computer Vision, vol. I, pp. 304-317, 2008.

[11] D. Yang and A. Śluzek, A low-dimensional local descriptor incorporating TPS warping for image matching, Image and Vision Computing, vol. 28(8), pp. 1184-1195, 2010.

[12] C. Schmidt and R. Mohr, Object recognition using local characterization and semi-local constraints, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19(5), pp. 530-534, 1997.

[13] M. Paradowski and A. Śluzek, Local Keypoints and Global Affine Geometry: Triangles and Ellipses for Image Fragment Matching, Innovations in Intelligent Image Analysis, SCI 339, pp. 195-224, 2010.