Machine Vision and Applications manuscript No.
(will be inserted by the editor)

When Standard RANSAC is Not Enough

Cross-Media Visual Matching with Hypothesis Relevancy*

Tal Hassner · Liav Assif · Lior Wolf

Received: date / Accepted: date

Abstract The same scene can be depicted by multiple visual-media. For example, the same event can be captured by a comic image or a movie frame; the same object can be represented by a photograph or by a 3D computer graphics model. In order to extract the visual analogies that are at the heart of cross media analysis, spatial matching is required. This matching is commonly achieved by extracting key-points, and scoring multiple, randomly generated mapping hypotheses. The more consensus a hypothesis can draw, the higher its score.

In this paper we go beyond the conventional set-size measure for the quality of a match and present a more general hypothesis score that attempts to reflect how likely is each hypothesized transformation to be the correct one for the matching task at hand. This is achieved by considering additional, contextual cues for the relevance of a hypothesized transformation. This context changes from one matching task to another and reflects different properties of the match, beyond the size of a consensus set. We demonstrate that by learning how to correctly score each hypothesis based on these features we are able to deal much more robustly with the challenges required to allow cross media analysis, leading to correct matches where conventional methods fail.

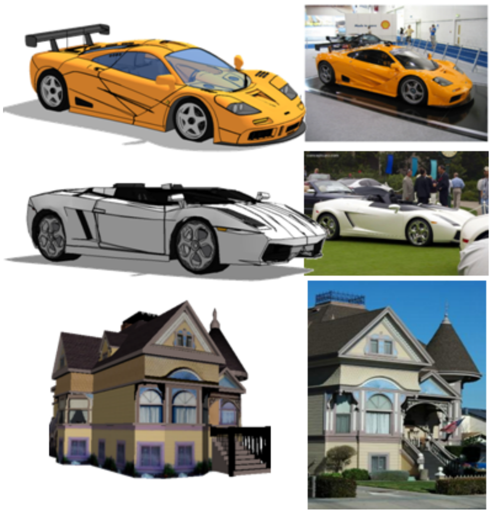


Fig. 1 Example application: matching a 3D model (left) to a photograph (right). The viewpoint of the 3D object was adjusted to match the recovered viewpoint of the photograph.

Keywords Object detection · 3D viewpoint estimation · Image registration · Shape matching

1 Introduction

Being able to accurately match different representations of the same visual scene is a key enabling requirement in many computer vision and graphics systems

T. Hassner
Department of Mathematics and Computer Science, The Open University of Israel, Israel
1 University Road, P.O.B. 808, Raanana 43107, Israel
Tel.: +972-9-778-1261
Fax: +972-9-778-0605
E-mail: hassner@openu.ac.il

L. Assif
Department of Mathematics and Computer Science, The Open University of Israel, Israel

L. Wolf
Blavatnik School of Computer Science at Tel Aviv University, Israel

* The final publication is available at springer.com, DOI: 10.1007/s00138-013-0571-4

(e.g., [1,2]). Matching often takes the form of recovering the parametric transformation relating these representations. This can be a homography, aligning different photos of the same scene, a projection matrix describing how a 3D shape projects onto a photo, and more.

Solutions to such problems are well known in the computer vision community. They typically assume that features are extracted in each representation and then matched. The parametric transformation is then estimated by solving an overdetermined system of equations (e.g., by Least Squares). In most cases, however, many of these feature correspondences are erroneous. A robust estimation procedure, such as the Random Sample Consensus (RANSAC) algorithm [3], is therefore used to obtain an estimate in the presence of "outlying" correspondences.

RANSAC works by randomly selecting a small subset of putative correspondences and using these to produce a hypothesis of the transformation's parameters. The remaining correspondences are then examined to determine which agree with the transformation. In a traditional RANSAC implementation, this number, referred to as the size of the "consensus", or inlier, set, is taken as a measure of the quality of the estimate. A large number of hypotheses are generated and ranked. The hypothesis with the largest inlier set is selected as the output transformation.

Although RANSAC has proven highly successful in matching different images from the same source (e.g., when producing panoramic photos [4]), we show here that it is far less successful when different sources of visual information are considered. Such cases are particularly challenging for representation and matching techniques and so often provide only small percents of correct putative matches. This, in turn, requires standard RANSAC implementations to perform what can easily become unacceptable numbers of iterations [4]. Moreover, in many situations, outliers might be counted as inliers and the selected transformation would not be the best transformation among the random hypotheses.

Our key observation in this paper is that by applying standard RANSAC, without explicitly considering the underlying problem it is employed to solve, we are blinding ourselves to important problem-specific clues for the quality of hypotheses. We show that such clues may be used to obtain better transformations. Specifically, we consider the following three types of problem specific information in order to obtain more robust *hypothesis relevancy* measures. (i) **Inlier-set distributions.** The size of the inlier set and how its members are spatially distributed. (ii) **Appearance similarities.** Similarities measured between the transformed and the target photos. (iii) **Transformation likelihoods.** The likelihood of the recovered transformation and its components.

We show that these different measures of similarity may be combined into a single score, reflecting the relevancy of a hypothesis. In order to optimally combine these measures we suggest using statistical learning techniques, when training data is available, or domain knowledge when it is not. Our results demonstrate that by substituting the standard "max-inlier" measure of the quality of a hypothesis, with our hypothesis relevancy scores, we obtain far better transformations. This is verified both quantitatively and qualitatively on different matching tasks. Specifically, we present "pose estimation" results where photos are matched to Computer Generated images (CGIs) of digital, 3D models (Fig. 1), as well as affine transformations estimated between different representations of the same scene (cartoons or Lego figures matched with video frames and more).

2 Related work

RANSAC variants. Pose estimation and image alignment methods often use RANSAC [3] to find optimal transformation hypotheses. Over the years many variants of the original RANSAC procedure have been proposed and we only briefly touch on some related methods here. For a comprehensive survey we refer the reader to [5].

RANSAC extensions typically consider the inlier set alone in order to obtain a measure of the quality of a hypothesis; different techniques advocating different ways of extracting hypothesis scores from the inlier set and its spatial distribution [6,5]. In some cases, improved performance is obtained by better sampling strategies [7,8], pre-filtering of the set of correspondences [9,10], and faster computation of the parameters of each hypothesis [11]. Some methods attempt to tune RANSAC for real-time performance [12,13], while others focus on the quality of the final model [14,15] when applied to specific problems. Finally, RANSAC has also been shown to perform well for non-rigid alignment tasks in [16]. A comparative evaluation of some of these methods can be found in [12].

Recently, [17] proposed image similarity based measures of a hypothesis quality. Though somewhat related to our own approach, we consider multiple sources of information on the quality of hypotheses, and demonstrate how these may be combined in a manner which best suites the alignment task at hand.

Here, we propose a general approach which combines multiple measures of the quality of a hypothesis in order to suppress wrong hypotheses which gain high numbers of inliers, while promoting low-inlier hypotheses which provide acceptable solutions. To this end, we employ Statistical machine learning. Although such methods have been used before in conjunction with RANSAC (e.g.,RANSAC-SVM [18] and, more recently, [19,20]), these have used RANSAC to improve the quality of the machine learning models required for subsequent classification, whereas here, we use machine learning as a means for selecting better RANSAC estimates.

Image to image alignment. Correlation based direct methods have been proposed as a means of aligning different visual representation of the same scene, while overcoming their differing appearances [21]. When the scenes are non-rigid, or else present different interpretations of the same visual information, correspondence based methods are often more suitable. Matches established between key-points in the two images, provide a means for estimating the parameters of a desired transformation. Much of the attention of previous methods has focused on improving the repeatability of the key-point detectors [22], the robustness, descriptiveness, and compactness of the local representations [23], and the quality of the matching [24]. The work presented here augments these methods by focusing instead on how a particular parameter hypothesis is evaluated. It can therefore be applied along-side any of these techniques in order to provide better quality transformations.

Pose estimation. Numerous methods have been described for estimating the 6-degrees of freedom pose of a camera. Broadly, these can be categorized into two main groups: methods using image-based models for the underlying geometry of the object, and methods employing explicit, 3D representations.

A large number of photos may be used to capture the appearance of an object from different viewpoints and thus facilitate pose estimation. This approach has the advantage that typically it is easier to compare images of the same modalities rather than photos to CG images. The downside is the requirement of having multiple, often a great deal, of photos to capture the appearance of the object from all possible viewing angles [25–27].

Related to our work is the alternative approach of using explicit 3D information. 3D models have been exploited in different ways in the past, typically by using a CG representation of the object. A popular approach is to compute pose-estimation and segmentation jointly by using the object's contour. Some examples include [28,29]. Although contours often provide accurate information, they are sensitive to occlusions, they do not provide sufficient information when objects are smooth or convex, and they may be mislead by background noise. To improve accuracy, some methods propose making local features more robust to certain geometric transformations (e.g., [30]), however these do not provide solutions to matching between real and synthetic textures.

Texture information on the 3D geometry has directly been exploited by a number of existing methods. These form matches between an input photo and a rendered CGI view of the 3D model acting as a proxy for the 3D geometry [31–34]. More recently, this approach has been combined with recognition [35] and detection [36,37]. These methods use many 3D models from the same class, employing correspondences between query features and features from multiple CG views. All these methods use RANSAC to obtain the final pose. Here we augment these methods by considering multiple measures for the quality of each pose estimate.

3 Preliminaries - RANSAC

The RANSAC algorithm has been applied to many robust estimation tasks. Here, we consider it specifically for the purpose of computing the transformation from a source to a target image, where these images may be of different media types (e.g., different modalities). Specifically, an initial, global set $\mathcal{G}$ of putative correspondences is formed between key-points in the two images to be matched. RANSAC then operates by iterating the following two steps: hypothesis *generation* and hypothesis *verification*. In the first step, a set of correspondences $\mathcal{S} \subset \mathcal{G}$ is randomly selected and then used to estimate a *hypothesis* – the values for the parameters of the transformation from the source to the target image. The size of $\mathcal{S}$ is typically the smallest possible size from which a hypothesis may be extracted. For an Affine transformation relating two images, for example, three 2D point matches provide six equations for the six unknown parameters.

Following the hypothesis estimation step, the obtained hypothesis is then evaluated and scored. Here, the remaining correspondences in $\mathcal{G}$ are consulted to determine the number k of correspondences which support the current hypothesis: A correspondence is said to support a hypothesis when applying the parametric transformation to a source point brings it to within a pre-determined distance d from its corresponding target point. The number k, the *number of inliers*, is traditionally taken as the measure of quality for the hypothesis; all hypotheses are sorted by k and the one with the highest value, the *max-inlier* iteration, is then used to produce the output transformation by using all its inlying correspondences to recompute the transformation.

The number of times these two steps are performed is usually determined by the empirically estimated ratio of the correct vs. incorrect putative correspondences in $\mathcal{G}$. With fewer correct matches, more iterations are required in order to ensure a high enough probability that a subset $\mathcal{S}$, randomly selected, will contain only correct matches. This number can quickly become unreasonable when the percentage of correct matches is small, as is often the case when matching between images from different sources.

4 Matching with hypothesis relevancy

Our key observation is that selecting a hypothesis based solely on the number of inliers is often misleading and is by no means the only source of information we have for the quality of a hypothesis. Consider for example Fig. 2. Here, RANSAC is used to compute the six-degrees of freedom camera poses which match 2D photos to 3D models of the same objects (see Sec. 5). The graphs present RANSAC iterations, sorted by the size of the inlier set, k, for each hypothesis. In Fig. 2 (Top), the value for k, the number of inliers in each iteration, reaches its maximum value in several different iterations, some providing a correct hypothesis while others do not. In Fig. 2 (Bottom), on the other hand, a suitable hypothesis for the camera pose was obtained in an iteration which did not score the highest number of inliers.

In both these cases a correct hypothesis may possibly be found by fine-tuning the value of d, the threshold determining when a match in $\mathcal{G}$ is an inlier for the hypothesis. Doing so, however, is not trivial: setting this value too high (a liberal threshold) would produce many iterations which score the maximal number of inliers (as in the top example in Fig. 2), whereas setting it too low (a conservative threshold) may miss inlying correspondences and would therefore be more sensitive to noise (bottom example in Fig. 2).

Here, instead of relying exclusively on inlier set sizes we consider additional hypothesis quality features, specific for the problems being considered. Sec. 4.1 describes these features while Sec. 4.2 describes how they may be combined in order to produce an alternative hypothesis score – the *hypothesis relevancy* score.

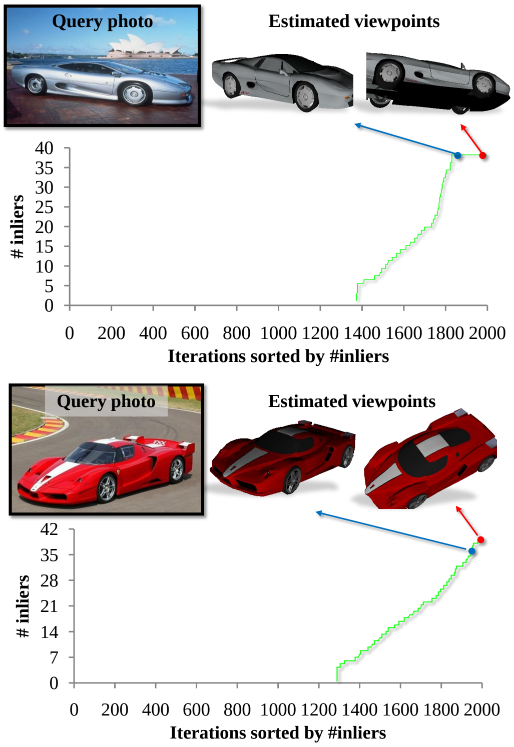


Fig. 2 RANSAC 2D to 3D matching based on maximum inliers alone. A query photo is matched to a 3D CG model of the same object by recovering the six degrees of freedom camera pose. **Top:** Several RANSAC hypotheses score the same number of inliers; some with a suitable pose (blue), others with a wrong pose (red). **Bottom:** The RANSAC hypothesis with the maximum number of inliers (red) is not the best hypothesis (blue).

4.1 Hypothesis relevancy features

We consider the following three types of *hypothesis relevancy features* for the quality of a hypothesis.

Inlier-set distributions. The number of inliers and the spatial distribution of these inliers provide important clues for the quality of a hypothesis. We therefore employ both the number of inliers (the traditional measure for the quality of a hypothesis) and the inlier convex-hull size (measured as a percent of the image size) as two hypothesis relevancy features. We expect a good hypothesis to include points spread out across much of the image, whereas a poor hypothesis to involve inliers concentrated in only a small area of the image. The higher the value of this second feature, the better the hypothesis is considered.

Appearance similarities. We consider the correspondences formed by matching descriptors extracted at key-points. These descriptors capture the visual information local to each key-point. We evaluate the similarities of these descriptors in each inlier set, seeking a hypothesis for which the descriptors in the source and

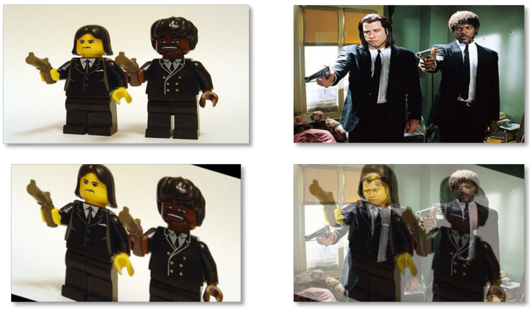


Fig. 3 Image matching with domain knowledge (Sec.4.2). Top row are the input source Lego image and the target photo. Bottom row is the result of applying the recovered Affine transformation to the source image (left) and overlaid on the target (right). By applying machine learning, better suited matches can be obtained (see Sec. 7.2).

target image have similar appearances. Specifically, we compute for each corresponding pair consistent with the scored hypotheses (i.e., each inlying correspondence) the sum of squared differences between its SIFT descriptors, obtaining a vector whose length is the number of inliers. From this vector of distances we derive five features, namely, their mean, SD, median, minimum, and maximum.

Transformation likelihoods. These features depend on the particular transformation we seek to recover; the features used for camera pose estimation (Sec. 5) are different from those used for image-to-image Affine matching (Sec. 6). In the former case, the features are based on the difference between the viewpoint angles of a synthesized view of the 3D model and the photograph. In the latter, they are based on the parameters of the aligning transformation. We detail these features at length in Sec. 5 and Sec. 6 respectively.

4.2 Combining quality measures

In the previous section we proposed a number of features which may be examined to provide a better picture of the quality of a hypothesis. The question now is how to combine these separate features in order to obtain a single hypothesis relevancy score?

Applying domain knowledge. When domain knowledge is available, indicating for each feature what values are associated with good hypotheses and which suggest bad ones, Fisher's combined probability test [38,39] can be used to merge the features into a single relevancy score. Specifically, we convert each feature score into an empirical p-value by taking its percentile out the values obtained for the same feature in all other RANSAC iterations. In other words, since the vast majority of the hypotheses are wrong, the distribution of the per-hypothesis score for a given feature closely matches the distribution under the null hypothesis that the hypothesis is false; the percentile of a given feature provides an estimate for the significance of its score. Note that domain knowledge is used here by determining which end of the distribution (high or low values) is desirable.

Combining multiple scores, the relevancy score for hypothesis j is then computed by:

$$\chi_j^2 = -2 \sum_{i=1}^{R} log_e(p_i) \qquad (1)$$

Where $i \in [1..R]$ is a feature index, with R features used for the current application. A result of this method, applied to the task of image-to-image matching, is presented in Fig. 3. We next explain how these results may be improved by applying machine learning techniques.

Learning to combine features. When domain knowledge is unavailable or insufficient, we instead use Statistical machine learning to obtain a weighing of the features into a single relevancy score. We collect a training set consisting of image pairs, representing instances of the matching problem at hand. We obtain ground truth estimates for the desired transformations linking the members of each couple. We then use RANSAC to obtain feature values for all iterations. We compute the value of each feature i across all training iterations, and use these to linearly normalize the feature values to the range of [0..1].

Having the ground truth transformations at our disposal, we attach each iteration with a positive / negative label of whether it provided an acceptable hypothesis, or not (see Sec. 5 and Sec. 6 for details on this process for particular matching tasks). We then train a discriminative classifier on the feature vectors, using these labels as targets. In all our tests we used the simple and parameter-free Linear Discriminative Analysis (LDA) classification algorithm [40]. Once trained, a hypothesis is scored by projecting its feature vector onto the 1D LDA subspace. Although better performance may presumably be obtained with more sophisticated classifiers, we focus in this work on informative features rather than on optimizing classification engine.

5 3D model matching

We consider both matching of 3D models to photos and cross-media photo matching. Naturally, the first task is more involved, and we therefore describe it first. The adjustments needed to match between cross-media 2D photographs will be described in Sec. 6, based on the more elaborate system.

Given a 3D, CG model m and a photo, I^m, of the same object, taken with a camera whose unknown external parameters are given by some rotation matrix R and translation vector t, we wish to recover the six degrees of freedom of these parameters in the CG model's coordinate frame, thus matching the 3D model and the photograph.

Having m at our disposal allows us to render images of the model, producing CG image (CGI) views V_j^m. Each view includes, besides its intensities, also the 3D coordinates of the points projected onto each of its pixels. By establishing a link between a pixel x_i in I^m and pixel x'_i in V_j^m, we obtain the pairs (x_i, X_i), were X_i is the 3D point m's coordinate frame, projected onto x'_i. These can then be used to estimate the viewpoint of I^m using standard camera calibration methods [4]. Specifically, given pairs (x_i, X_i), the matrix $R_{3\times3}$ and vector $t_{3\times1}$ may be obtained by solving

$$x_i \sim \ A[R \ \ t]X_i \tag{2}$$

Where $A_{3\times3}$ is the intrinsic camera matrix, and R is constrained to be an orthonormal matrix.

For simplicity, we assume that the focal length is known and set it to 800 in image pixels, that the principle point is at the image center, that the pixel's aspect ratio is one, and that there is no skew. Our method is agnostic to the type of camera calibration model that is used to estimate the pose from point matches and it is straightforward to relax these assumptions by using more elaborate camera calibration techniques.

We obtain image to image correspondences by computing standard feature descriptors, here, the SIFT descriptors [41], on Harris-Affine detected points [42]. Each descriptor in I^m is matched to its L2-nearest neighbor in V_j^m. Pose can then be recovered by employing RANSAC as a robust estimator [3]. Though it conceivably possible to improve the performance of the system by the use of more robust features (e.g., [43]), or prior knowledge [44], we focus here on the pose-estimation process rather than improving the quality of the point matches.

If multiple CGIs V_j^m exist and a sufficient number of correct matches is established in each of these views, then this process should yield the same pose estimate for all views. In practice, however, as previously mentioned, the overwhelming presence of many false matches results in pose estimates that vary greatly between the different CGIs. We therefore choose the hypothesis which obtained the highest score from amongst all views as the final, output transformation.

5.1 Learning pose hypothesis relevancy

We next detail how the relevancy of pose hypotheses can be learned from training data. Here, we assume a training set of a certain class of 3D CG models and associated photos of these objects. In Section 7.1 we show that our system is robust to the selection of these models, photos and their class. The camera poses for the photos included in this training set are computed by manually establishing point correspondences between the photos and the CGIs.

For every *training* model m we render CGI views V_j^m, covering the object's viewing sphere. For each of its training images, I_k^m we then estimate the pose automatically using each one of these rendered views, separately. Each such estimate provides us with (i) a pose error e_{jk}^m computed by comparing the hypothesis with the ground truth pose and (ii) the features characterizing the quality of the pose estimate. These features are collected in a vector v_{jk}^m (Sec. 4.1), one for each view.

For every CGI view V_j^m in the training set, and for every given photo I_k^m of the same model m we obtain an estimate of the pose in I_k^m. This is then compared to the (known) ground-truth pose and an error is computed as a function of the angular and translational difference between the estimated pose and the ground truth pose. This error serves to compute training labels used to produce the learned, LDA model. The pair (I_k^m, V_j^m) is assigned a label of 1 if the error e_{jk}^m falls below a predefined threshold and -1 otherwise. In other words, the positive class is the class of feature vectors computed for the relevant views.

In our implementation we define e_{jk}^m based solely on angular differences. It is measured as the angle between the principle axes of the known and estimated positions of the bounding box of m. Let $p \in \mathbb{P}^3$ be a point on m (in homogeneous notation), $\hat{T}$ defined as

$$\hat{T} = \begin{bmatrix} \hat{R} & \hat{t} \\ \mathbf{0} & 1 \end{bmatrix}$$

be the estimated extrinsic matrix, and T similarly defined using the ground truth rotation R and translation t. Assuming a fixed camera matrix, we compute:

$$\hat{p} = \hat{T}T^{-1}p \tag{3}$$

Points $\hat{p}$ are then used to produce the estimated bounding box and compute e_{jk}^m.

The feature vectors v_{jk}^m, along with the labels computed based on pose estimation accuracy, are used to train a discriminative model for selecting relevant views.

The LDA classifier obtained is used to link the features extracted from new CGIs of novel models to novel photos. During the application (test) phase, the feature vector v'_j is computed as above for each hypothesis in each CGI view. The LDA classifier is then employed on these vectors to obtain a numeric score that is expected to be positive and high if the hypothesis is accurate and negative and low otherwise. This numeric score is used to rank the hypotheses and identify the most relevant one. This is repeated for all views. The highest scoring hypothesis, across all views, is selected as the final, output hypothesis.

5.2 Specific features for 3D-to-2D matching

Beyond the relevancy features described in Sec. 4.1, we employ *Transformation Likelihood* relevancy features for the particular task of 3D-to-2D matching. To this end, we consider the correspondences established between the query photo interest points and interest points in each rendered CGI view V_j^m (Sec. 5). Each such set of correspondences yields a pose estimate for the query photo (see Eq. 3). This pose – the position of the query photo relative to the 3D model – can be compared to the known, automatically-specified pose of the current rendered view V_j^m, and the angle between the two, α_{kj}^m, can be determined (See illustration in Figure 4). In practice, α_{kj}^m is computed similarly to e_{jk}^m (Eq. 3) using the known extrinsic matrix of CGI view V_j^m and the estimated matrix of I_k^m.

We observe that a large value for the angle α_{kj}^m between the estimated pose and the reference view V_j^m can be due to an actual large difference in poses. But this is unlikely, as if this was indeed the case, the rendered view and the query photo will likely appear different and so few correspondences, if any, will be accurate. More likely is that such a large angle resulted from false correspondences and an erroneous pose estimate. Small differences, on the other hand, are either the result of a correct estimate (i.e., the query photo was taken from a pose close to that of the rendered view), or, again, an unreliable estimate. Assuming a uniform distribution of erroneous estimates, however, it is less likely for a small angle difference to be the result of an error.

As we report in Sec. 7.1, this feature proved to provide the most influence on the hypothesis relevancy score computation. We note that an alternative approach of manually limiting the range of admissible pose estimates for query I^m and rendered reference V_j^m to be smaller than some angle $\hat{\alpha}_{kj}^m$. Beyond the disadvantage of having to specify these values manually, and possibly manipulating them for different objects and object classes, this has the additional adverse effect of imposing a hard, single threshold on all the views. This, compared to the soft, learned values computed for each object class and weighed against other features to determine more informative hypothesis relevancy scores.

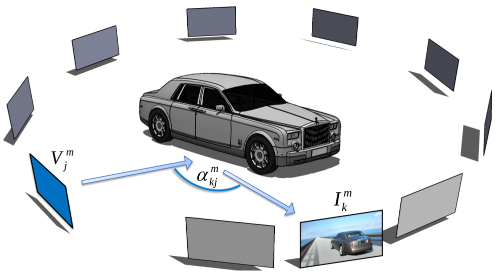


Fig. 4 Photo-to-CGI pose difference. Illustrating the angle α_{kj}^m between the pose estimate for photo I_k^m using matches between its image-features and the CG view's V_j^m.

6 Matching cross-media photos

As an additional example of cross-modality matching, we consider the task of obtaining an Affine transformation between different media capturing the same visual scene. As in the pose estimation problem above, the task here can be particularly challenging when an exact transformation does not exist due to the differing representations. This is made more challenging by the representations themselves having different appearance properties, leading to a reduced probability of forming correct correspondences.

We treat the 2D-to-2D matching task similarly to that of matching CGIs and photos (Sec. 5); the latter viewed as a particular instance of the former. Of course, unlike the pose estimation task, we have only a single "view". As a parametric model we use Affine transformations, which are powerful enough for our purposes, yet require fewer parameters than full projective transformations. Each hypothesis stems from three randomly selected correspondences and is scored based on a learned hypothesis relevancy score, using the features described in Sec. 4.1.

To obtain suitable Transformation Likelihood features, we employ QR-decomposition in order to extract the translation and scale along the X and Y axes, as well as the shear value and rotation angle from the affine transformation matrix. These six parameters are used as features based on the assumption that the probabilities of the possible Affine transformations are not uniform; some Affine transformations are more likely than others given the task at hand.

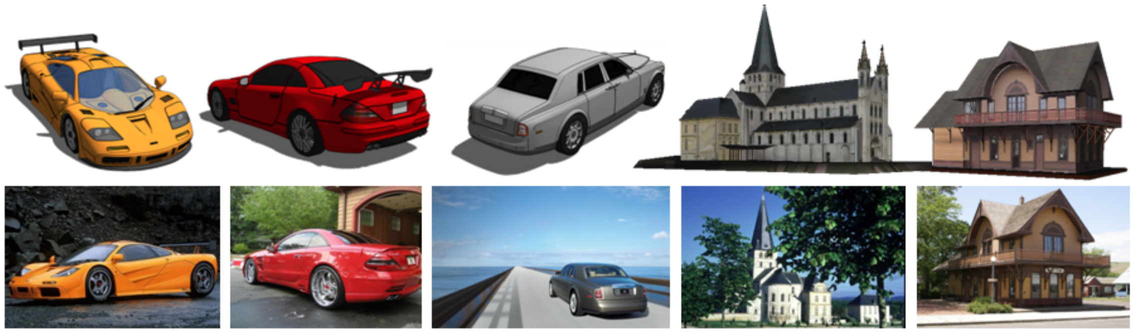


Fig. 5 Example query+model pairs. Top row are rendered views of 3D CG models, from arbitrary viewpoints. Bottom are example query photos collected from the web.

7 Experiments

We present results in multiple cross-media domains. Quantitative experiments focus on the matching of 3D views to photographs, since it is easy to define a meaningful error in such cases. Qualitative experiments are presented for various additional cross media domains, such as computer games to real-world, Lego models to movie frames, and comics to motion pictures, in order to demonstrate the applicability of our method, even in extreme, cross-media matching tasks.

Our method is implemented in MATLAB, using a MATLAB OpenGL wrapper for rendering the CG models in our pose estimation tests. Standard OpenCV routines were used to compute transformations in both the 2D-3D and the 2D-2D experiments.

7.1 Quantitative experiments

Cars and Buildings benchmarks. We have assembled benchmark data sets and ground truth data suitable for evaluating 2D-to-3D matching. Specifically, we have collected textured, 3D, CG models of car and building objects, along with images from the web, taken of those same objects. Our models were obtained from the Google 3D Warehouse collection and the images were downloaded from Wikipedia. In total, we have 31 car models with 90 test images and 11 building models with 30 images, models having one to three query images each. All models were scaled to unit size. Car models were further roughly aligned – all facing the same direction. Finally, we recover the ground-truth camera pose of all our test images by establishing manual correspondence between the images and CGI views of their associated CG model. Fig. 5 presents some examples of our models and test images.

With this data set, we define a straightforward leave-one-out testing protocol, as follows. Given an image, we estimate the pose of the object in the image, using the object's CG model. In addition, all other models, their images, and ground truth poses are available for training; the only excluded information is, of course, the ground truth pose of the input image, as well as all other query photos of the same object. Pose estimate precision is measured following [37] by considering both the translational ϵ_t and angular ϵ_r errors. Specifically, ϵ_t is the difference between the center of the ground truth model and the center of the model in the estimated position, ϵ_r is the angle between the principle axes of the real and estimated bounding boxes [37]. We use Eq. 3 to obtain the estimated position of the object's bounding box.

Comparison with existing work. We compare our method to the RANSAC-based method of [37]. We note that better pose estimation accuracy may conceivably be obtained by more recent systems (e.g., [45–47]). We build on the system proposed in [37], however, as it allows us to focus on the contribution of our modified RANSAC routine, rather than those of other components of a 3D, pose estimation system (e.g., descriptor design and matching, etc.)

Similarly to [37], for every model m we produce 324 CG views V_j^m: 108 views uniformly covering the upper hemisphere of the object at three radii. Descriptors are extracted using the Harris-Affine interest point detector implementation of [42]. SIFT descriptors were computed using the code made available by [48]. Given a descriptor set extracted from a novel photo we match each descriptor against those of the current CGI view seeking its nearest neighbor in Euclidean distance.

Pose is then estimated using $2,000$ RANSAC iterations using these putative correspondences. In training, hypotheses which produce angular errors of 7 degrees or

Please see the project webpage for available resources, including our MATLAB functions for rendering and computing the transformations. URL: www.openu.ac.il/home/hassner/projects/ransaclearn

Source: sketchup.google.com/3dwarehouse

less are considered positive samples, all others are considered negative samples. When LDA is applied, the hypothesis with the highest LDA projection value is selected, and its pose estimate is then returned as our method's output.

Cars and Buildings benchmark results. Table 1 compare the performance of the following methods on the Cars and Buildings benchmarks:

1. **Random view + RANSAC.** A CGI view V_j^m is selected randomly and its matches are then used to estimate the pose using standard RANSAC (Sec. 3).
2. **Best view + RANSAC.** The view selected for pose estimation is the one with the most nearest neighbor matches for the descriptors in the query photo. Once selected, pose is estimated as before.
3. **Estimation based on [37].** This method is used as a baseline to our own, in order to evaluate the effect of the modified RANSAC procedure. In order to remain true to their implementation, however, we perform testing using all our training models, including the model of the object appearing in the test photo.
4. **Our learned hypothesis relevancy method.** The method described in Sec. 5.
5. **Robustness to training.** Same as 4, but here car estimates were produced using a statistical model learned from the Buildings set and vice-versa.

Table 1 summarizes results for both Cars and Buildings sets, listing angle and position median and mean ± standard error (SE). The angular error of our method outperforms other variations by significant margins. Position errors, on the other hand, vary little from one method to the other, all doing well. This is unsurprising considering that translation can be estimated, to a large degree, based on a crude key-point localization within an object's boundaries, which all methods do well. Figure 6 demonstrates this point; the type of errors obtained in the rotational model by the method of [37] have little effect on the location of the object.

It is interesting to consider the weights learned for the different features involved in computing the hypothesis relevancy scores, presented in Figure 7. Apparently, the most influential feature is the Photo-to-CGI pose difference (Sec. 5.2). The traditional feature used for selecting hypotheses – the inlier set size – is second in its influence on the hypothesis scores, but with a substantially smaller contribution. This implies that using standard RANSAC, without learned domain knowledge, may lead to sub-optimal accuracy. The same is evident by comparing the contribution of the inlier set size feature to other features – particularly the maximum similarity of inlier descriptors and the size of the convex hull – which appear to be nearly as important.

To evaluate the generality of our learned LDA classifiers, we performed an additional experiment (Table 1, row 5). Here, an LDA classifier trained on our Cars set was used to compute hypothesis scores when estimating the poses of objects from the Buildings set, and vice versa. For the car objects, the drop in performance was minor, with median and average angular errors increasing only slightly. For the building objects, however, results actually improved. By offering more examples in more variable viewing positions, the Cars set provides a richer and more effective training set, thereby improving accuracy on the Buildings set. Where traditional techniques may be unable to learn from examples having fundamentally different appearances, here, by focusing on the properties of the *transformations*, rather than the object *appearances*, we can utilize training across object classes.

Stability of results. We analyze the stability of our approach by measuring pose estimation accuracy with different numbers of RANSAC iterations. Figure 8 plots the median angular errors obtained for varying numbers of iterations, using our full approach – the learned hypothesis relevancy method (row 4 in Table 1). Evidently, pose errors quickly drop and remain stable from around 700 iterations onwards. These should be considered along with the cross-dataset training results (row 5 in Table 1), as testament to the robustness of our approach.

The limitations of our methods are presented in Figure 9. The method is challenged by similarity among completely different views or by lack of details in the given photo. While our criterion for hypothesis selection improves performance, the problem of multiple hypotheses testing may still lead to the identification of wrong transformations.

Car detection by matching 3D models. We evaluate our method on the cars in the image set from [49] testing for detection accuracy using the Pascal VOC07 evaluation protocol [50] and 8-class, pose classification accuracy. We use the same estimated 3D car model computed by [25] as our reference model m, and compare to their detection and pose classification results.

We have successfully detected 114 cars out of 160 (71.25%) compared to the 61.25% of [25]. Fig. 10 compares the diagonal of the confusion matrix of the two methods relative to the ground truth labeling of the eight pose labels. Detection based on matching with hypothesis relevancy outperformed [25] on all but one class. We note that better performance on this bench-

	Cars				Buildings			
	Angular Error		Position Error		Angular Error		Pos. Error	
Method	Median	Mean ± SE	Median	Mean ± SE	Median	Mean ± SE	Median	Mean ± SE
1. Rnd. view + RANSAC	117.9	116.6 ± 5.49	1.73	1.86 ± 0.29	91.04	90.56 ± 7.53	0.89	1.36 ± 0.31
2. Best view + RANSAC	93.09	96.32 ± 5.30	1.19	2.01 ± 0.53	77.53	79.92 ± 8.43	0.84	1.30 ± 0.27
3. Liebelt et al. [37]	66.47	77.52 ± 5.74	0.58	**0.83 ± 0.11**	48.40	58.19 ± 7.67	0.55	0.81 ± 0.19
4. Hyp. rel. score	**18.55**	**42.01 ± 5.14**	**0.10**	0.94 ± 0.32	22.26	39.78 ± 7.41	**0.08**	**0.66 ± 0.28**
5. Hyp. rel., cross-class	19.10	45.68 ± 5.49	**0.10**	1.02 ± 0.33	**21.66**	**36.30 ± 6.81**	0.11	**0.66 ± 0.28**

Table 1 Precision statistics. Median and mean (± standard error of the means, SE) angular and position errors on the Cars and Buildings data sets for all tested methods. Lower values are better.

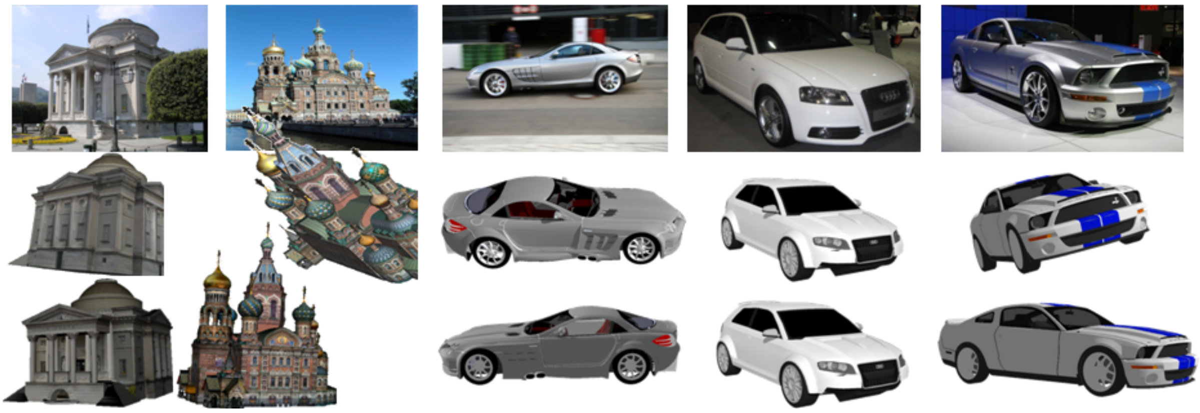


Fig. 6 Visually comparing pose estimates of our method to [37]. Top row is the input photo, middle is [37] and bottom our results. Note that the churches in the second column were moved in the figure to allow a more compact presentation; the translation component is accurate for both methods.

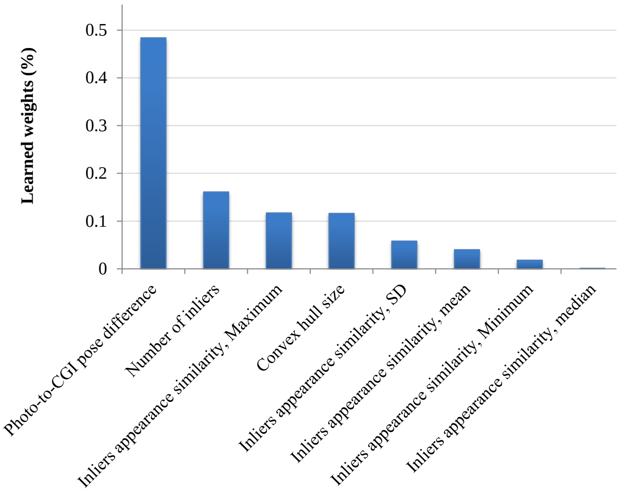


Fig. 7 Learned weights for features used to compute our hypothesis relevancy scores. See text for more details.

mark have recently acheived by [26]. Their method, however, uses a far more accurate reference 3D model which was unavailable to us for comparison.

7.2 Qualitative experiments

We performed additional cross media domain experiments, focusing on 2D image to image matching tasks. Given an image pair, we seek the Affine transformation linking the two images. Here, we again extract SIFT key-points, this time, however, we use every fifth pixel along edges detected by the Canny edge detector as key-points. This, in order to obtain a sufficient number of key-points even in low contrast images (e.g., comics in Fig. 11(a)). Training in all these examples is performed in a leave-one-out manner, similar to our 2D-to-3D matching experiments (Sec. 7.1).

We present results of matching comics to frames from the motion picture “300” in Fig. 11(a), matching of Lego models to photos of the same scenes in Fig. 11(b), and screen-shots of the video game MineCraft to photos of similar figures in Fig. 11(a). In all cases training was performed using similar example data (e.g., pairs of comics and frames from “300” were used to train an LDA model for matching other comics to frames taken from the same motion picture). No additional parameter tuning was performed, and we used the same features in all these experiments (Sec. 4.1).

Source: www.minecraft.net

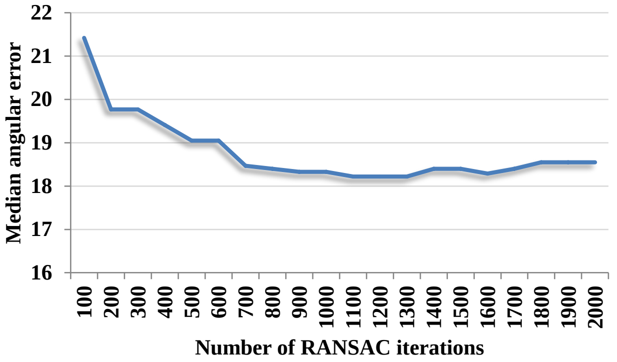


Fig. 8 Stability tests on the Cars benchmark. Median angular errors on the Cars benchmark measured for our full approach (row 4 in Table 1) with increasing numbers of RANSAC iterations.

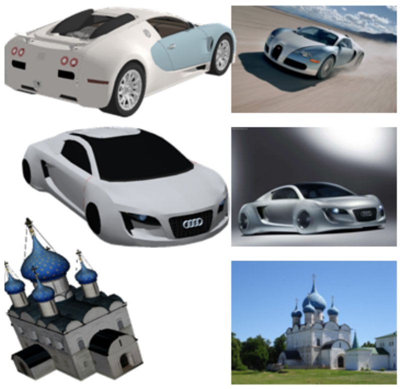


Fig. 9 Examples of failed estimations. These are typically cases where the object appears similar from different views (top), has few features (middle), or are caused by poor random hypothesis selection by RANSAC (bottom).

We compare hypothesis relevancy to the RANSAC-based, Gold-Standard algorithm for aligning images [4], in both cases using the same key-points and descriptors. Here, again, more elaborate alignment schemes exist (e.g., the recent work of [51]), but our goal is to evaluate the performance of the modified RANSAC, rather than fine-tune an alignment pipeline.

As can be seen, in some cases, (e.g., Fig. 11(a) and (c), last rows) hypothesis relevancy and max-inliers both obtain similar transformations. In most cases, however, using hypothesis relevancy instead of max-inliers greatly improves the quality of the obtained alignment. A difficult example where both methods failed is presented in the last row of Fig. 11(b).

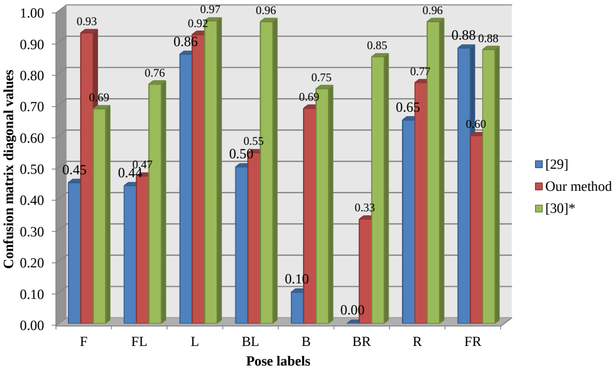


Fig. 10 Pose classification. Comparing our confusion matrix diagonal values to [25] on the benchmark in [49]. Higher values are better. * We show results also for [26] though we note that they used a substantially better 3D model, which was not available to us in our experiments. These results are therefore not directly comparable to our own.

8 Conclusions

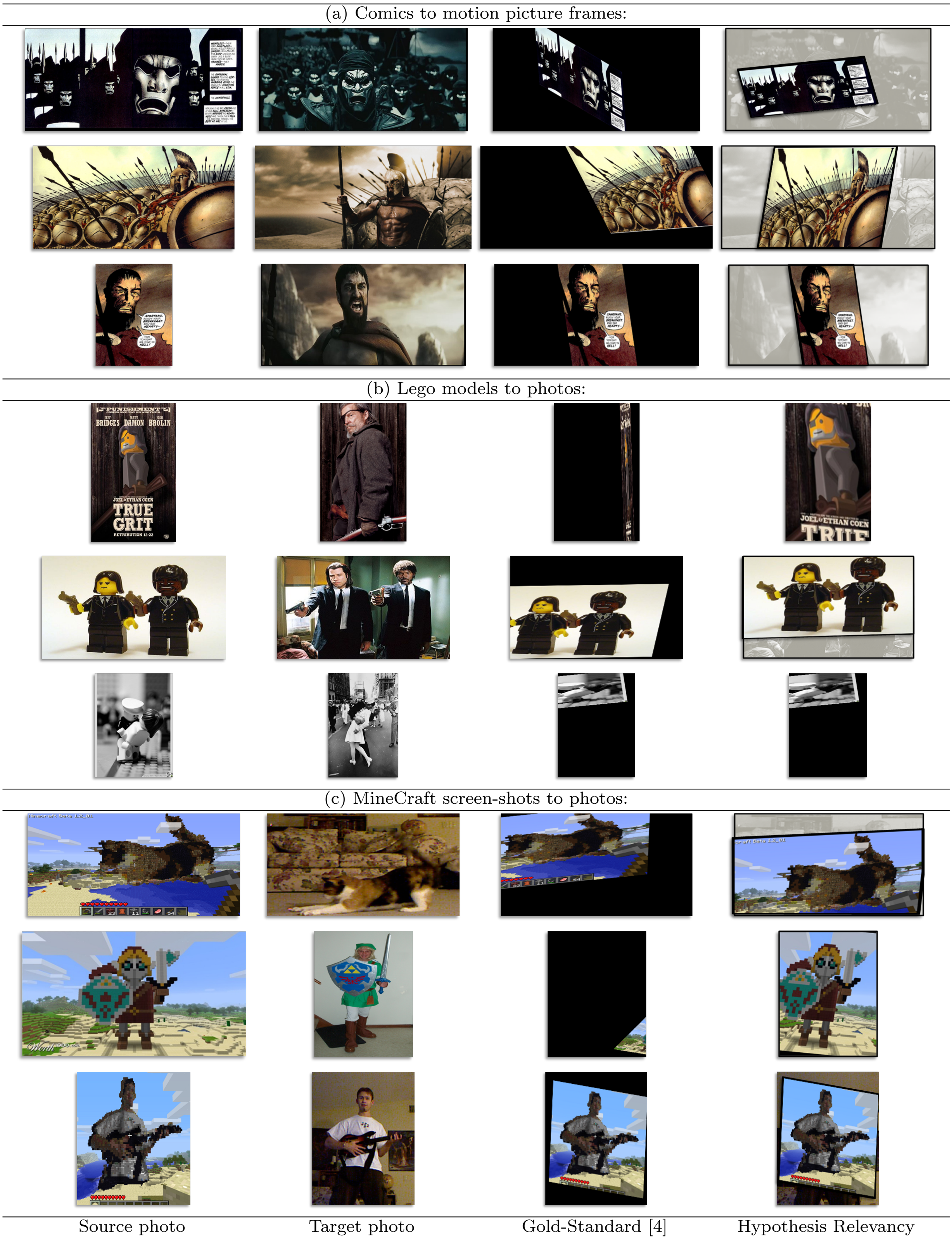
Matching across different modalities is a challenging task that results in a potentially large number of false matches. Furthermore, it is not easy to distinguish between true and false matches even when considering consensus among multiple matches. Conventional tools such as RANSAC often fail to identify sets of matches that support a correct hypothesis from sets that support false hypotheses that have equally high or even better scores due to a nasty combination of inaccurate matches and multiple hypothesis testing.

In this work we propose to augment the RANSAC procedure by considering multiple sources of information, combined using a learning based relevancy score. This has the effect of making the RANSAC procedure far more robust. Overall, the simplicity of our method makes the proposed solution practical and efficient, and quantitative results on three benchmarks, as well as a variety of qualitative results, demonstrate its effectiveness. In addition, multiple qualitative experiments in various cross-media applications demonstrate its utility.

Acknowledgements TH was partially funded by General Motors (GM).

References

1. Cui, X., Kim, H., Park, E., Choi, H.: Robust and accurate pattern matching in fuzzy space for fiducial mark alignment. MVA (2012) 1–13



(a) Comics to motion picture frames:

(b) Lego models to photos:

(c) MineCraft screen-shots to photos:

Fig. 11 Qualitative results, matching different 2D image representations of the same scenes. Affine transformation estimated from a source image to its target image. Warped source images shown for both the Gold-Standard method [4] using RANSAC with max-inliers as well as hypothesis relevancy (shown here overlaid on the target images).

2. Yoon, S., Scherer, M., Schreck, T., Kuijper, A.: Sketch-based 3D model retrieval using diffusion tensor fields of suggestive contours. In: ACM-MM, ACM (2010) 193–200
3. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. Com. of the ACM **24** (1981)
4. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press, ISBN: 0521540518 (2004)
5. Choi, S., Kim, T., Yu, W.: Performance evaluation of RANSAC family. In: BMVC. (2009)
6. Capel, D.: An effective bail-out test for RANSAC consensus scoring. In: BMVC. (2005) 629–638
7. Chum, O., Matas, J.: Matching with PROSAC-progressive sample consensus. In: CVPR. Volume 1. (2005) 220–226
8. Matas, J., Chum, O.: Randomized RANSAC with sequential probability ratio test. In: ICCV. Volume 2., IEEE (2005) 1727–1732
9. Chin, T., Yu, J., Suter, D.: Accelerated hypothesis generation for multi-structure data via preference analysis. TPAMI (2012) 1–1
10. Sattler, T., Leibe, B., Kobbelt, L.: SCRAMSAC: Improving RANSAC's efficiency with a spatial consistency filter. In: ICCV, IEEE (2009) 2090–2097
11. Botterill, T., Mills, S., Green, R.: Fast RANSAC hypothesis generation for essential matrix estimation. In: Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on, IEEE (2011) 561–566
12. Raguram, R., Frahm, J., Pollefeys, M.: A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. ECCV (2008) 500–513
13. Scaramuzza, D.: Performance evaluation of 1-point-RANSAC visual odometry. JFR **28** (2011) 792–811
14. Frahm, J., Pollefeys, M.: RANSAC for (quasi-) degenerate data (QDEGSAC). In: CVPR. Volume 1., IEEE (2006) 453–460
15. Torr, P., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. CVIU **78** (2000) 138–156
16. Tran, Q.H., Chin, T.J., Carneiro, G., Brown, M., Suter, D.: In defence of RANSAC for outlier rejection in deformable registration. ECCV (2012) 274–287
17. Yan, Q., Xu, Y., Yang, X.: A robust homography estimation method based on keypoint consensus and appearance similarity. In: ICME, IEEE (2012) 586–591
18. Nishida, K., Kurita, T.: RANSAC-SVM for large-scale datasets. In: ICPR, IEEE (2008)
19. Bozkurt, E., Erzin, E., Erdem, Ç., Erdem, A.: RANSAC-based training data selection for speaker state recognition. In: InterSpeech. (2011)
20. Nishida, K., Fujiki, J., Kurita, T.: Multiple random subset-kernel learning. In: CAIP, Springer (2011) 343–350
21. Ukrainitz, Y., Irani, M.: Aligning sequences and actions by maximizing space-time correlations. ECCV (2006) 538–550
22. Aanæs, H., Dahl, A., Steenstrup Pedersen, K.: Interesting interest points. IJCV (2011) 1–18
23. Chandrasekhar, V., Takacs, G., Chen, D., Tsai, S., Reznik, Y., Grzeszczuk, R., Girod, B.: Compressed histogram of gradients: A low-bitrate descriptor. IJCV (2012) 1–16
24. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. TPAMI **27** (2005) 1615 – 1630
25. Arie-Nachimson, M., Basri, R.: Constructing implicit 3D shape models for pose estimation. In: ICCV. (2009)
26. Glasner, D., Galun, M., Alpert, S., Basri, R., Shakhnarovich, G.: Viewpoint-aware object detection and pose estimation. In: ICCV, IEEE (2011)
27. Su, H., Sun, M., Fei-Fei, L., Savarese, S.: Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In: ICCV, IEEE (2009) 213–220
28. Prisacariu, V., Reid, I.: PWP3D: Real-time segmentation and tracking of 3D objects. In: BMVC. (2009)
29. Sandhu, R., Dambreville, S., Yezzi, A., Tannenbaum, A.: Non-rigid 2D-3D pose estimation and 2D image segmentation. In: CVPR. (2009) 786–793
30. Wu, C., Clipp, B., Li, X., Frahm, J., Pollefeys, M.: 3D model matching with viewpoint-invariant patches (VIP). In: CVPR. (2008) 1–8
31. Gall, J., Rosenhahn, B., Seidel, H.: Robust pose estimation with 3D textured models. Advances in Image and Video Technology (2006) 84–95
32. Hassner, T., Basri, R.: Example based 3D reconstruction from single 2D images. In: Beyond Patches Workshop at CVPR. (2006)
33. Hassner, T., Basri, R.: Single view depth estimation from examples. CoRR **abs/1304.3915** (2013)
34. Hassner, T.: Viewing real-world faces in 3D. In: ICCV. (2013)
35. Stark, M., Goesele, M., Schiele, B.: Back to the future: Learning shape models from 3D CAD data. In: BMVC. (2010)
36. Liebelt, J., Schmid, C.: Multi-view object class detection with a 3D geometric model. In: CVPR. (2010) 1688–1695
37. Liebelt, J., Schmid, C., Schertler, K.: Viewpoint-independent object class detection using 3D feature maps. In: CVPR. (2008)
38. Fisher, S.: Statistical methods for research workers. Number 5. Genesis Publishing Pvt Ltd (1932)
39. Whitlock, M.: Combining probability from independent tests: the weighted z-method is superior to fisher's approach. Journal of evolutionary biology **18** (2005) 1368–1373
40. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. TPAMI **19** (1997) 711–720
41. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
42. Mikolajcyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV **60** (2004) 63–86 http://www.robots.ox.ac.uk/ vgg/research/affine/.
43. Hassner, T., Mayzels, V., Zelnik-Manor, L.: On sifts and their scales. In: CVPR, IEEE (2012) 1522–1528
44. Van Kaick, O., Tagliasacchi, A., Sidi, O., Zhang, H., Cohen-Or, D., Wolf, L., Hamarneh, G.: Prior knowledge for part correspondence. Computer Graphics Forum **30** (2011) 553–562
45. Gu, H.Z., Lee, S.Y.: Car model recognition by utilizing symmetric property to overcome severe pose variation. MVA (2012) 1–20
46. Hu, W.: Learning 3D object templates by hierarchical quantization of geometry and appearance spaces. In: CVPR, IEEE (2012) 2336–2343
47. Xiang, Y., Savarese, S.: Estimating the aspect layout of object categories. In: CVPR, IEEE (2012) 3410–3417
48. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/ (2008)

49. Savarese, S., Fei-Fei, L.: 3D generic object categorization, localization and pose estimation. In: ICCV. (2007)
50. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge 2007 (VOC2007) results. www.pascal-network.org/challenges/VOC/voc2007 (2007)
51. Lin, W.Y., Liu, L., Matsushita, Y., Low, K.L., Liu, S.: Aligning images in the wild. In: CVPR, IEEE (2012) 1–8