17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

On the Influence of Viewpoint Change for Metric Learning

Marco Filax and Frank Ortmeier
Chair of Software Engineering, Otto-von-Guericke University, Germany
{firstname.lastname}@ovgu.de

Abstract

Physical objects imaged through a camera change their visual representation based on various factors, e.g., illumination, occlusion, or viewpoint changes. Thus, it is the inevitable goal in computer vision systems to use mathematical representations of these objects robust to various changes and yet sufficient to determine even minor differences to distinguish objects. However, finding these powerful representations is challenging if the amount of data is limited, such as in few-shot learning problems. In this work, we investigate the influence of viewpoint changes in modern recognition systems in the context of metric learning problems, in which fine-grained differences differentiate objects based on their learned numeric representation. Our results demonstrate that restricting the degrees of freedom, especially by fixing the virtual viewpoint using synthetic frontal views, elevates the overall performance. We await that our observation of an increased performance using rectified patches is persistent and reproducible in other scenarios.

1 Introduction

A variety of works [1, 11, 20, 21] investigated the influence of viewpoint changes in the context of man-made features like SIFT [10]. It was shown that unwarping of perspective distortion introduced through synthetic viewpoint changes enhance the overall matching performance because a SIFT feature embeds information of edges in the context of the image plane.

With the rise of deep learning methods, the influence of viewpoint change was abandoned due to the astonishing baseline of learned classifiers [13]. This is because it is sufficient to include enough samples of a given class from different viewpoints to achieve viewpoint invariance for the mid-level features computed within a CNN before their final classification output.

Unfortunately, it is not always possible to include enough samples per class with different viewpoints. This is especially challenging for few-shot learning problems. An example of this type of problem is recognizing grocery products. A large variety of different grocery products are available that can only be distinguished by subtle visual differences [4]. While it is indeed possible to gather a database that comprises multiple samples per product from different viewpoints for every product currently available, it is considered time-consuming, error-prone, and costly. It is, therefore, common to use the available images from the web for products that typically depict a product from only a single (or just a few) viewpoint. These are usually taken under studio conditions and serve the purpose of presenting the benefits of the product to potential customers. Thus, only a small number of shots are available for products to train a neural net.

In this work, we investigate the influence of synthetic viewpoint changes in the context of grocery product recognition. We synthesize the frontal views of every grocery product patch for a given database. The database of grocery products was collected with a spatial aware camera and therefore comprises all required information to compute a synthetic frontal patch of every product of interest. This work's primary contribution is the sound evidence of a performance gain through restricting degrees of freedom during the imaging of an object in the context of a metric learning problem.

The remainder of this work is structured as follows. In the following, we present related works to the aspect of grocery recognition and viewpoint synthesis. We describe the underlying idea and our approach in Section 3. In Section 4, we report on our experiments with synthetic fronto-parallel views. We thereby trained multiple models and rigorously evaluated them. We conclude our work in Section 5.

2 Related Work

The influence of viewpoint changes for artificial features like SIFT [10] has been studied in a variety of other works, e.g., [1, 11, 20, 21] to name a few. ASIFT [11, 20] is the most influential work in this field. The core idea is to sample multiple embeddings per SIFT feature by sampling different viewpoints across the view hemisphere. It is proven to work reasonably well under the assumption that the object imaged by a camera is almost planar.

The influence of viewpoint changes has been studied in at least two works in the context of metric learning. [17] investigates the influence of *small* viewpoint changes on the problem of face recognition. Thereby the authors propose to sample synthetic views with a change of 5°, 10°, or 20° between viewing and testing. [16] synthesized a complete dataset for person re-identification. Thereby the authors proposed to move synthetic avatars in a vertically placed circle w.r.t. the camera. This work aims to evaluate the influence

of substantial viewpoint changes for a real-world fine-grained dataset.

Grocery recognition can be solved as a metric learning problem [4, 18, 19]. These works' core idea is to learn an embedding function that maps crops of images of stock-keeping units into an embedding space. Thereby the authors gained the ability to distinguish unseen products at test time. However, the impact of viewpoint change has not been studied in this field to the best of our knowledge.

3 Concept

Distinguishing groceries is a challenging problem due to the fine-grained visual differences that separate products from another. Further, the number of available product classes, i.e., individual stock-keeping units, grows continuously. These aspects render classification systems to be unfeasible in practice [4]. Recently, other approaches were used to solve these problems efficiently. Their overall idea is to predict an embedding based on the visual appearance of the product. These embeddings are then used to group similar products, i.e., the same SKUs, based on their Euclidean distance.

From an architectural perspective, models trained to solve that problem are similar to standard classification models, except for the last final layers such that the training goal differs. These metric learning goals have proven to be well suited in the domain of face recognition [14], in which similar problems arise. The substantial influence of dramatic viewpoint changes under real-world constraints was not addressed in detail in the related works (cf. Section 2). However, some works have already evaluated the influence of *small* viewpoint changes [17] or completely synthetic datasets [16].

In this work, we evaluate the influence of dramatic viewpoint changes w.r.t. stability of embedding under another real-world scenario: the fine-grained recognition of grocery products. This is mainly because grocery product recognition is a few-shot metric learning problem that relies on only a few database samples from an almost frontal viewpoint similar to the faces used for face recognition.

3.1 Fronto-Parallel Synthetic Views

We unwarp the perspective transformation of grocery product candidates using the well-known homography [5]. Thereby we generate a fronto-parallel view of a skewed bounding box of a stock-keeping unit such that it is rectangular, just as in the real world under studio conditions. We compute a homography H that maps the quadrilateral bounding box to a rectangular shape, w.r.t. the real-world skewed bounding box, which is typically not axis aligned. Finally, we warp the source image with this homography. A set of 2D image point and their corresponding set of 2D rectangular

Figure 1. Some examples from the database. The first column depicts a stock-keeping unit from the web. The second column depicts examples from shelves. Both taken from [3]. The last column depicts the proposed synthetic frontal view.

points, mathematically defined as four correspondences $x_i \leftrightarrow x'_i$, we solve the following equation

$$\mathbf{x'_i} \times H\mathbf{x_i} = 0. \tag{1}$$

Whereas $\mathbf{x_i} = \{x_i, y_i, w_i\}$ and $\mathbf{x'_i} = \{x'_i, y'_i, w'_i\}$ are homogeneous 2D points and H is a 3x3 matrix. The equation is solved using the direct linear transformation algorithm [5].

3.2 Recognizing Grocery Products

To solve the problem of grocery recognition, we train an *embedding network* with a particular training goal, which is described by the *loss function*. An embedding network can be described as a function $f_\theta(x)$ such that $f_\theta(x) : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^D$ whereas $f_\theta(x)$ is parameterized by θ and transforms an image from $\mathbb{R}^{n \times n}$ to a point on the manifold $\mathbb{R}^D$. The core idea is to map images of the same SKU to metrically close points on $\mathbb{R}^D$. Similarly, $f_\theta(x)$ shall transform images of different SKUs to distant points.

We adopt a solution proposed for face recognition[2] and use the well-known ResNet-50 [6] as the base network. We remove the final classification head and the average pooling layer. The architecture is then as follows: a maximum pooling layer, a batch normalization

layer [8], followed by a dropout layer [15], and a fully connected embedding layer, which is followed by another batch normalization layer [8].

We deploy an online *triplet loss function* based on [4] to learn the embeddings of images by minimizing the Euclidean distance of the anchor images x_a and an image depicting the same stock-keeping unit x_p. Similarly, we maximize the distance between the anchor x_a and an image x_n of another grocery product. The loss $\mathcal{L}(\theta, \mathcal{B})$ is defined as

$$\mathcal{L}(\theta, \mathcal{B}) = \sum_{i=1}^{Y} \sum_{a=1}^{K} [\log(1 + \exp(\max_{p=1..K} (D_{f_\theta}(x_a^i, x_p^i)) - \min_{\substack{j=1..Y \\ n=1..K \\ i \neq j}} (D_{f_\theta}(x_a^i, x_n^j))))]. \tag{2}$$

Thereby D_{f_θ} is defined as $D_{f_\theta}(x^i, x^j) = ||f_\theta(x^i) - f_\theta(x^j)||_2^2$. $\mathcal{B}$ describes a batch of images and Y is the set of classes. Following [7], we use the hardest positive and hardest negative pairs of $\mathcal{B}$ to emulate moderate triplets. K is the number of samples to be drawn for every class to account for possible outliers.

4 Experiments

In this section, we report on our experiments to evaluate the impact of viewpoint unwarping in the context of grocery product recognition.

We use a real-world grocery product dataset [3] to conduct our experiments. The dataset has in total over 20.000 different grocery products. 871 of these stock-keeping units have examples from within stores which were extracted from 41.955 camera images. The bounding boxes of objects were semi-automatically labeled on a camera stream. The semi-automatically creation of the dataset was based on Microsofts HoloLens. Thus, the trajectory of the camera is known for all of these images taken in the store. Thereby the 2D quadrilateral bounding box for every annotated product on the shelves is also known. This allows us to unwarp the perspective distortion introduced with viewpoints changes of the camera. We generate a synthetic fronto-parallel view for every annotation using planar homographies as described in Section 3.1. Thereby we unwarp the skewed bounding boxes into squares. We compute the set of points X' for every 2D point $x \in X$ of the skewed quadrilateral bounding box in image coordinates to define the complete set of corresponding points. We calculate the homography H that maps points in X to the points X' for every bounding box based on these four correspondences. Via applying these, we generate a frontal view for every annotation.

We conduct our experiments with the skewed original image crops of grocery products - which are identical to the crops published in [3] and our fonto-parallel synthetic crops derived from the original crops. Thus, the number of grocery products and the number of examples per class are completely identical. Practically, the only thing that differs for both experiments is the camera's (synthetic) viewpoints during imaging. We, therefore, evaluate the influence of viewpoint changes in uncontrolled settings, such as grocery stores.

Table 1. Experimental results for a grocery recognition task. All stock-keeping units were unknown at test time. We trained multiple models in a cross folded manner with different embedding dimensions in two settings: using skewed original crops and their synthesized fronto-parallel views. We depict the mean results of all cross-folds. The best results are highlighted in bold.

	Embedding Dimension			
	512	256	128	64
Skewed	65.6%	66.8%	62.8%	**61.9%**
Frontal	**68.3%**	**67.6%**	**64.9%**	61.1%

We train two different embedding functions which aim at detecting the fine-grained visual differences of grocery product crops. The hyperparameters in both settings are identical. As a base network, we chose the well-known ResNet [6] and initialized the weights with ImageNet weights. We cut the final classification layer and replaced the previous averaging pooling layer with a global max-pooling layer. Afterward, we deployed a BN-Dropout-FC-BN structure for the embedding part, similar to [2]. The dropout rate is fixed at 0.6. We employ standard forms of augmentations for all images, such as scaling, shifting, and adding noise.

The embedding functions are trained with Adam [9], a batch size of 170, and a learning rate of 5×10^{-4} without decay. We trained the models for 2000 epochs to ensure maximal performance in both cases. We deploy a rigorous evaluation protocol [4] to address existing flaws [12] in other metric learning protocols. We split the dataset into three complete disjoint sets w.r.t. their product classes. That means that there is no overlap in terms of SKUs during training, validation, or testing. The classes, i.e., SKUs are disjoint. We preserve 171 SKUs for evaluation and use the remaining 700 SKUs in a three cross-fold validation to train three different embedding functions. Note that the SKUs used for training and evaluation are identical for both experiments with skewed and fronto-parallel views. To evaluate the performance of the trained embedding functions, we report the mean recall@1.

Table 1 summarizes the results of our experiments. We repeated the proposed experiment with different embedding dimensions to grasp the performance across different problem difficulties. Our results show that the mean performance of the embedding functions trained with skewed examples is better than models trained

with frontal image patches. As all hyperparameters in the experiments were fixed during training, we conclude that the performance gain can only be due to the fronto-parallel view synthesis. The gain in performance, especially if the model is powerful enough to solve the task reasonably well (cf. higher embedding dimensions), is due to the task's reduced complexity because some degrees of freedom in the imaging of objects are eliminated through the frontal view synthesis. We, therefore, conclude that the use of frontal image patches for metric learning boosts performance.

5 Conclusion

In this work, we evaluated the influence of viewpoint changes in the context of metric learning for the practical example of grocery recognition. The key idea is based on the observation that the viewpoint introduces additional degrees of freedom during an object's imaging. Using synthetic views that assemble a fronto-parallel projection of a product's underlying view plane within a shelf, we were able to demonstrate a performance gain using a rigorous state-of-the-art evaluation protocol. We demonstrated that using synthetic frontal views boosts the overall performance by fixing all necessary hyperparameters in our experiments, which we executed cross-folded. In the future, we want to investigate these findings on other databases and integrate the presented approach into a mobile app.

References

[1] Cai, G. R. et al. "Perspective-SIFT: An efficient tool for low-altitude remote sensing image registration". In: *Signal Processing* 93.11 (2013), pp. 3088–3110.

[2] Deng, J. et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *arXiv:1801.07698* (2018).

[3] Filax, M., Gonschorek, T., and Ortmeier, F. "Data for Image Recognition Tasks: An Efficient Tool for Fine-Grained Annotations". In: *ICPRAM*. SciTePress, 2019, pp. 900–907.

[4] Filax, M., Gonschorek, T., and Ortmeier, F. "Grocery Recognition in the Wild: A New Mining Strategy for Metric Learning". In: *VISAPP*. SciTePress, 2021, pp. 498–505.

[5] Hartley, R. and Zisserman, A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004, pp. 1–646.

[6] He, K. et al. "Deep Residual Learning for Image Recognition". In: *CVPR*. IEEE, 2016, pp. 770–778.

[7] Hermans, A., Beyer, L., and Leibe, B. "In Defense of the Triplet Loss for Person Re-Identification". In: *arXiv:1703.07737* (2017).

[8] Ioffe, S. and Szegedy, C. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv:1502.03167* (2015).

[9] Kingma, D. P. and Ba, J. L. "Adam: A method for stochastic optimization". In: *arXiv:1412.6980* (2019).

[10] Lowe, D. G. "Distinctive Image Features from Scale-Invariant Keypoints". In: *Int. J. Comput. Vis.* 60.2 (2004), pp. 91–110.

[11] Morel, J.-M. and Yu, G. "ASIFT: A New Framework for Fully Affine Invariant Image Comparison". In: *SIAM J. Imaging Sci.* 2.2 (2009), pp. 438–469.

[12] Musgrave, K., Belongie, S., and Lim, S.-N. "A Metric Learning Reality Check". In: *arXiv:2003.08505* (2020).

[13] Razavian, A. S. et al. "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition". In: *CVPRW*. IEEE, 2014, pp. 512–519.

[14] Schroff, F., Kalenichenko, D., and Philbin, J. "FaceNet: A unified embedding for face recognition and clustering". In: *CVPR*. IEEE, 2015, pp. 815–823.

[15] Srivastava, N. et al. "Dropout: A simple way to prevent neural networks from overfitting". In: *JMLR* 15 (2014), pp. 1929–1958.

[16] Sun, X. and Zheng, L. "Dissecting Person Re-Identification From the Viewpoint of Viewpoint". In: *CVPR*. IEEE, 2019, pp. 608–617.

[17] Swystun, A. G. and Logan, A. J. "Quantifying the effect of viewpoint changes on sensitivity to face identity". In: *Vision Res.* 165 (2019), pp. 1–12.

[18] Tonioni, A. and Di Stefano, L. "Domain invariant hierarchical embedding for grocery products recognition". In: *Comput. Vis. Image Underst.* 182 (2019), pp. 81–92.

[19] Tonioni, A., Serra, E., and Di Stefano, L. "A deep learning pipeline for product recognition on store shelves". In: *IPAS*. IEEE, 2018, pp. 25–31.

[20] Yu, G. and Morel, J. "A fully affine invariant image comparison method". In: *Int. Conf. Acoust. Speech Signal Process.* 26.1 (2009), pp. 1597–1600.

[21] Zhang, Z. et al. "TILT: Transform invariant low-rank textures". In: *Int. J. Comput. Vis.* 99.1 (2012), pp. 1–24.