Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Decomposition of Multimodal Data for Affordance-based Identification of Potential Grasps

Daniel Dornbusch, Robert Haschke, Stefan Menzel, Heiko Wersing

2012

Preprint:

This is an accepted article published in 1st Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM 2012). The final authenticated version is available online at: https://doi.org/[DOI not available]

Decomposition of Multimodal Data for Affordance-based Identification of Potential Grasps

Daniel Dornbusch[1], Robert Haschke[1], Stefan Menzel[2] and Heiko Wersing[2]
[1]*CoR-Lab, Bielefeld University, Bielefeld, Germany* [2]*Honda Research Institute Europe GmbH, Offenbach, Germany*
ddornbus@cor-lab.uni-bielefeld.de

Keywords: Decomposition Algorithms : Multimodal Data : Grasp Identification

Abstract: In this paper, we apply standard decomposition approaches to the problem of finding local correlations in multi-modal and high-dimensional grasping data, particularly to correlate the local shape of cup-like objects to their associated local grasp configurations. We compare the capability of several decomposition methods to establish these task-relevant, inter-modal correlations and indicate how they can be exploited to find potential contact points and hand postures for novel, though similar, objects.

1 INTRODUCTION

We tackle the challenging problem of finding suitable grasps for unknown objects employing well-known decomposition approaches. In order to grasp an unknown object, previously acquired grasping knowledge from similar objects needs to be exploited and adapted to the current situation. Because similar object shapes, e.g. handles, also afford similar hand postures for grasping, we are looking for correlations between different modalities of successful grasping examples. Based on depth images and object silhouettes extracted from color images, we aim for a prediction of contact locations as well as an associated hand posture to realize the grasp. In doing so, we especially focus on local shape features, because different local parts of an object afford different grasps.

Unsupervised decomposition algorithms are able to find statistically relevant correlations in high-dimensional data sets, and thus are well-suited to the task at hand. Applying them to a multimodal data set allows for the identification of inter-modal, semantically meaningful correlations. The simultaneous use of multimodal data in decomposition approaches can: (i) improve the interpretability of the extracted basis components of each single modality, and (ii) extract functionally relevant correlations between different modalities. In Sec. 3, several decomposition methods, briefly introduced in the next section, are compared w.r.t. their ability to establish relevant inter-modal correlations in training data, and to infer grasps for new, though similar, objects. Finally, in Sec. 4 we discuss the results and draw some conclusions.

2 MATRIX DECOMPOSITION

The starting point of all decomposition approaches is a set of L vectors $\mathbf{x}_i \in \mathbb{R}^M$ pooled in an input matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$. Each $\mathbf{x}_i$ can be regarded as an observation of M random variables, comprising several modalities. We aim for a more compact, approximate representation of these observations using a small set of $N < M \ll L$ meaningful components spanning a new vector space $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_N]$. This assumes that observations lie on a low-dimensional manifold of $\mathbb{R}^M$, which here is approximated by a linear subspace. The basis vectors $\mathbf{f}_i$ will express typical correlations within the training set, also including correlations between different modalities of the data. Expressing the data vectors $\mathbf{x}_i$ with respect to the basis $\mathbf{F}$ yields an approximation matrix, $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_L]$. The $N \times L$ matrix $\mathbf{G}$ of coefficients is known as the encoding matrix.

Formally, we can restate this approach as matrix factorization, $\mathbf{X} \approx \mathbf{F} \cdot \mathbf{G} \equiv \mathbf{R}$, which minimizes the reconstruction error between the original data, $\mathbf{X}$, and its factorization, $\mathbf{F} \cdot \mathbf{G}$. The computation of $\mathbf{F}$ and $\mathbf{G}$ depends on the actual decomposition approach, which may impose additional constraints on both matrices, e.g., sparseness or non-negativity.

For example, PCA (Zhao et al., 2008) computes basis vectors that are pairwise orthogonal and point in the directions of the largest variances. k-Means clustering (Li and Ding, 2006) represents observations by a set of prototypes, $\mathbf{f}_i$, resulting in an extremely sparse encoding: only the coefficient associated to the nearest prototype equals one. Non-negative Matrix Factorization (NMF) (Zhao et al., 2008) constrains basis

vectors and encodings to non-negative values to avoid cancellations of features and facilitates their interpretability. Non-negative Matrix Factorization with Sparseness Constraints (NMF-SC) (Hoyer, 2004) is based on NMF and additionally enforces sparseness on the encodings and/or the basis components.

3 APPLICATION TO GRASPING

We compare the presented decomposition approaches in a grasping scenario to investigate their ability to find local inter-modal correlations. Based on a dataset of successful grasps applied to a set of cups, a compact set of basis components is calculated. In a subsequent application step, partial observations are augmented by a reconstruction of the missing modalities. To this end, encodings are computed based on existing modalities and missing ones are predicted from the corresponding linear combination of basis components. Finally, the best grasp can be chosen and realized by a robot hand.

3.1 Capturing of Grasping Data

To gather multimodal information of human grasping processes, the Manual Interaction Lab was created at CITEC, Bielefeld (Maycock et al., 2010). For the work presented in this paper, data from three modalities were captured: hand postures (motion-tracking coordinates), color video images and depth images. 16 different cup-like objects were selected to record grasping sequences belonging to three different grasp types: cup grasped by handle, from above, or from the side. 413 grasp configurations were captured, comprising 8-9 grasps per object and grasp type.

3.2 Preprocessing of Grasping Data

The captured raw sensor data was synchronized and preprocessed to obtain suitable input data for the grasp selection task.

Visual modalities. The grasp for a particular object is first and foremost determined by the shape of the object. A preliminary study using color images for decomposition resulted in basis components dominated by colors and textures. Hence, we decided to extract the object silhouette from these images, i.e. those pixels constituting the object shape. We also removed constant background pixels from all color and depth images, replacing them with zero values. Thus, the decomposition approaches do not need to explicitly model these irrelevant image parts. Contact regions on the object silhouette were identified by comparing images before and after establishing the grasp. All depth and color images were centered, cropped to the foreground region, and resized for normalization purposes. The image sizes of the sparse input modalities were 144×100 for the object silhouettes and contact regions and 61×46 for the depth images.

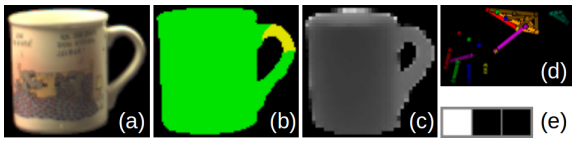
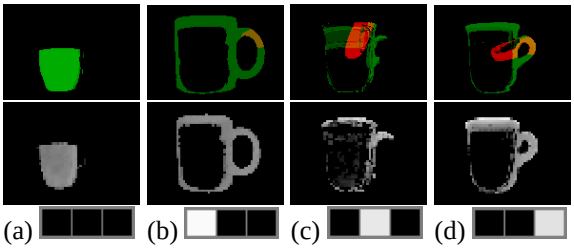


Figure 1: Modalities: (a) Color video image. (b) Object silhouette and contact areas. (c) Swiss Ranger depth image. (d) Visualization of Vicon coordinates. (e) Grasp type.

Hand posture. Hand posture sequences, obtained from tracking markers on all finger segments and subsequent calculation of the associated hand posture (Maycock et al., 2011), can be utilized in two manners: using the whole grasping trajectory or the final grasp posture only. In preliminary studies, we found that complete trajectories can be reconstructed in many cases. However, different grasping speeds and large variations of hand trajectories prior to actual grasping sometimes lead to visible dilatation effects in the reconstructed trajectories. Dynamic Time Warping (Mühlig et al., 2009) could compensate for asynchronous execution speeds and might in future work allow direct "replay" on a robot hand. In this paper, only the final hand pose is considered, adding a 27×3 dimensional vector of marker positions to the input data.

Grasp type. To distinguish the three employed grasp types, we could learn three individual sets of basis components, $\mathbf{F}_i$, employing appropriate subsets of the training data. However, this strongly reduces the number of data samples available for decomposition. Alternatively, a single decomposition could be applied to the entire training set comprising all grasp types, which often leads to an interference of basis components corresponding to different grasp types.

In order to choose a particular grasp type, we augmented all input vectors by an additional modality, employing three-dimensional unit vectors to indicate the grasp type. Then, we can explicitly request a particular grasp type by providing the corresponding unit vector as an additional input to the search process. This prevents simultaneous activation of basis components belonging to different grasp types, thus reducing co-activation of ambiguous local grasps. Furthermore, grasp-specific correlations between modalities are automatically labeled by the decomposition algorithms. Finally, the combined decomposition of all grasp configurations has the advantage that correlations, which are common to different grasp types,



(a) (b) (c) (d)

Figure 2: 4 of 250 exemplary basis components calculated by NMF-SC. Top: overlap of object silhouettes (green) and contact areas (red), middle: depth image, bottom: grasp type. Grasp posture not visualized. (a) Cup corpus. (b) Grasp by handle. (c) Grasp from above. (d) Grasp from side.

have to be learned just once.

All employed input modalities are summarized in Figure 1 along with the original color image, which is not included in the input data. Altogether, the data vectors comprise $14400 + 14400 + 2806 + 81 + 3 = 31690$ dimensions.

3.3 Calculation of Basis Components

In the first study, we applied the decomposition algorithms to a training data set consisting of 320 grasp configurations, derived from 13 of the 16 cup-like objects. We extracted subspaces spanned by 50, 100, 150, 200, 250 and 320 basis components. Additionally, we used a disjoint test data set comprising 93 grasp configurations belonging to the three remaining cups to measure the ability of the methods to generalize and represent novel, though similar, data. In this study, the encodings used for reconstruction were computed based on all modalities (depth images, hand postures, object silhouettes, contact areas and labeled grasp types). k-Means clustering produced prototypical, holistic and non-negative basis vectors. The encodings were maximally sparse and the ability to generalize was limited to the selection of the nearest cluster centroid. The principal components of PCA accomplished very small reconstruction errors, but were difficult to interpret due to their holistic nature and the occurrence of positive and negative values. The non-negative basis components produced by NMF were mostly sparse. The encoding matrix was always sparse and non-negative. NMF-SC computed only sparse, non-negative basis components and encodings such as the sparseness constraint was explicitly enforced by the algorithm. Figure 2 depicts four exemplary basis vectors, which represent local features like the cup handle or the body. As Table 1 shows, the normalized mean squared reconstruction error

$$\text{NMSE} = \frac{1}{L}\sum_{i=1}^{L}\frac{\|\mathbf{x}_i - \mathbf{r}_i\|^2}{\|\mathbf{x}_i\|^2}, \qquad (1)$$

Table 1: Study 1: Normalized mean squared reconstruction errors (NMSE) for observed (a) and test data set (b) against number of basis components. The encodings used for reconstruction are computed based on all modalities (depth images, grasp postures, object silhouettes, contact areas and labeled grasp types). Values scaled up by a factor of 10^3.

Approach	(a) Training Data					
	50	100	150	200	250	320
K-Means	75.7	49.4	31.7	18.3	10.8	0.0
PCA	39.1	20.4	11.7	6.4	2.8	0.0
NMF	52.1	31.1	20.2	12.1	7.8	2.5
NMF-SC	55.9	32.5	20.4	11.5	7.1	1.8
	(b) Test Data					
K-Means	163.2	167.6	165.5	176.0	174.1	179.0
PCA	91.7	80.3	74.2	70.8	68.0	65.6
NMF	107.1	97.8	94.7	95.1	94.9	98.2
NMF-SC	111.0	104.3	102.3	102.2	99.1	102.4

calculated on the *training data* decreases for all methods uniformly, as the number of available basis components is increased from 50 to 320. Using the maximal number of basis components, k-Means clustering and PCA are able to represent the training data perfectly. In contrast to this, for the *test data set* the inverse correlation between NMSE and the feature count did not always hold true. In particular, k-Means did not benefit from higher numbers of basis components. Also, NMSE increased again using the full set of basis vectors generated by NMF and NMF-SC due to overspecialization. Only PCA profited in all cases as additional basis component were added.

3.4 Identification of Potential Grasps for Novel Cup-like Objects

In the second study, we reconstructed missing modalities based on the inherent correlations between the basis components' subparts leading to the generation of potential grasp configurations for novel, though similar, objects. We calculated the encodings for the test observations based only on the following modalities: depth images (measured), object silhouettes (computed from color video images) and desired grasp types (set manually). In a second step, we utilized these encodings to reconstruct all modalities, including the previously neglected contact areas and grasp postures. Since the basis vectors were specialized to represent the correlations in grasping data, the missing information was approximated successfully for most novel objects (see Figure 3). Acknowledging the problem of overspecialization detected in the first study, we decided to investigate only the 150 (best error per component ratio) and 250 basis component (best before overspecialization) cases. To analyze the reconstruction errors for the approximation of

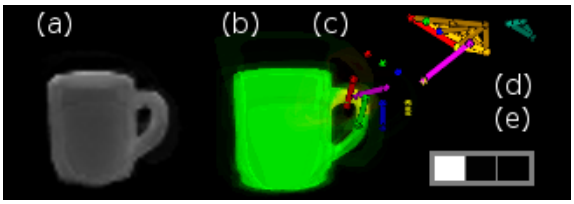


Figure 3: Potential grasp identified by NMF-SC for a new test cup, which was not used for training: Reconstructed depth image (a), contact area (c) and grasp posture (d) on top of the object silhouette (b), desired grasp type (e).

the missing modalities, we calculated the normalized mean squared reconstruction errors (NMSE) combined for all modalities (see Table 2a) and the combined NMSE only for the measured modalities (see Table 2b). Additionally, we evaluated the reconstructions of the formerly missing contact areas and grasp postures to assess the quality of the identified potential grasp configurations. This was done by calculating the mean contact region error E_m (see Table 2c), which is the average distance of the estimated and the expected contact area centroids (averaged over all L observations). Furthermore, we computed the mean fingertip error E_t (see Table 2d), which is the average error of the estimated and the expected fingertip positions. We found that the performance of k-Means on the test data set derogated using 250 feature vectors compared to 150, but still inferred the best overall grasp configurations for novel objects. PCA also produced good reconstruction errors, but sometimes resulted in diffuse contact areas for test objects due to its holistic basis features. NMF did not have this problem, because of its mostly sparse basis components. NMF-SC performed good at representing the measured modalities and the contact areas. However, the reconstruction error of generated grasp postures is larger. In summary, all algorithms were able to predict realistic grasp configurations in most cases.

3.5 Selection of Best Grasp

To select the best suited grasp for a novel object, a method has to be found in future work that is able to sort potential grasps by their quality, e.g., by evaluating the reconstruction errors of the measured modalities, or the sparseness of the inferred contact areas.

4 CONCLUSIONS

The decomposition approach provides a method to predict missing modalities from a few measurable modalities as long as strong correlations exist between both groups. We have shown that this approach can be used to predict hand configurations and desired contact regions for grasping an object based on its depth image and silhouette. This grasp information can subsequently be utilized for autonomous grasping – either by directly actuating a robot hand towards the estimated hand posture, or by computation of a hand posture realizing the estimated contact locations on the object. In both cases, an inverse hand kinematic can be used to obtain joint angles to actually operate the robot hand (Maycock et al., 2011).

Table 2: Study 2 – modality reconstruction: Encodings are computed based only on measured modalities and used to reconstruct missing modalities, i.e. contact areas and grasp postures. NMSE of (a) all and (b) measured modalities. NMSE scaled up by a factor of 10^3. E_m: Mean distance of estimated and expected contact area centroids (c). E_t: Mean error of estimated and expected fingertip positions (d).

Approach	Training Data							
	(a) AM		(b) MM		(c) E_m		(d) E_t	
	150	250	150	250	150	250	150	250
K-Means	37.3	13.2	63.4	23.4	1.79	0.76	6.2	2.5
PCA	25.2	5.5	61.5	13.8	1.13	0.21	8.1	3.2
NMF	35.4	11.4	78.2	23.7	2.64	0.76	9.1	3.9
NMF-SC	25.7	8.8	48.2	17.0	1.50	0.56	13.6	4.7
	Test Data							
K-Means	195	228	293	354	5.3	5.8	21.8	23.1
PCA	162	158	340	337	8.7	7.8	27.3	25.5
NMF	207	161	455	309	13.2	10.2	31.6	25.9
NMF-SC	151	139	258	233	5.5	5.3	52.0	33.2

ACKNOWLEDGEMENTS

Daniel Dornbusch gratefully acknowledges the financial support from Honda Research Institute Europe.

REFERENCES

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *JMLR*, 5:1457–1469.

Li, T. and Ding, C. (2006). The relationships among various nonnegative matrix factorization methods for clustering. In *Proc. ICDM*, Washington.

Maycock, J., Dornbusch, D., Elbrechter, C., Haschke, R., Schack, T., and Ritter, H. J. (2010). Approaching manual intelligence. *KI - Künstliche Intelligenz*, 24.

Maycock, J., Steffen, J., Haschke, R., and Ritter, H. (2011). Robust tracking of human hand postures for robot teaching. In *Proc. IROS*, San Francisco.

Mühlig, M., Gienger, M., Hellbach, S., and Goerick, C. (2009). Task-level imitation learning using variance-based movement optimization. In *Proc. ICRA*. IEEE.

Zhao, L., Zhuang, G., and Xu, X. (2008). Facial expression recognition based on PCA and NMF. In *Proc. WCICA*, Chongqing, China.