Representing Relative Visual Attributes with a Reference-Point-Based Decision Model

Marc T. Law
University of Toronto

Paul Weng
Shanghai Jiao Tong University
University of Michigan-Shanghai Jiao Tong University Joint Institute

***Abstract*—In many artificial intelligence, machine learning and computer vision tasks, the weighted sum model is used to value objects and define an order over them. In this paper, we consider two decision criteria defined as the (Euclidean and more generally Mahalanobis-like) distance to a reference point and investigate how they relate to the weighted sum model. In particular, we show that the distance-based representations can be seen as a relaxation of the representation induced by the weighted sum and we provide a characterization of the latter model with the former models in the case of strict orders. To illustrate our point, we consider the context of relative visual attributes. Nonetheless, our results also apply to other domains. More specifically, we present how these reference-point-based representations can be learned from pairwise comparisons and how they can be exploited for classification. Our experimental results show that those two criteria yield a more precise representation of the relative ordering for some attributes and that combining the best representations for each attribute improves recognition performance.**

I. Introduction

Ranking is a common and fundamental task in many domains, such as artificial intelligence, machine learning or computer vision. In those domains, objects to be compared are generally described as vectors in a feature space. The usual approach then relies on comparing objects through the values they are given by a linear function. This is for instance the case with the popular method called linear ranking SVM [1] in information retrieval. Such a linear function is called a weighted sum criterion in decision theory. Its widespread use and success can be explained by some of its nice properties, such as its simplicity and interpretability.

However, this simple criterion is not always suitable as suggested by axiomatic work in decision theory [2], [3]. It is known for example that not all preorders can be described by this linear model and using it to represent a certain ordering implies that some implicit assumptions (*e.g.* see Sections II and III) are made. Therefore, using this decision model may be problematic when such an assumption does not hold and one cannot expect an exact or good representation of an ordering when those assumptions do not hold.

In this paper, we consider two alternative representations for ranking that are reference-point-based. A reference-point-based criterion is defined as the Euclidean (or more generally Mahalanobis-like) distance to a fixed reference point. The first alternative representation is specified by setting the reference point to be an ideal point, the second one sets the reference point to be an anti-ideal point. An object is ranked higher if it is closer to the ideal point in the first case, or further from the anti-ideal point in the second case. Those criteria could be used in many diverse domains instead of the weighted sum criterion; we focus in this paper on the relative attribute problem [4] in computer vision to illustrate our point and demonstrate the usefulness of these representations.

II. Related Work

A. Attributes

In this paper, we consider visual attributes [5] which are high-level descriptions of concepts in images widely used in the computer vision community (see [6] and related work section therein). While traditional visual recognition approaches map low-level image features directly to object category labels, some recent works have proposed to focus on visual attributes. Generally, attributes have human-designated names (*e.g.* "is natural" or "is smiling", see Fig. 1) and are then valuable tools to give a semantic meaning to objects or categories in various problems. They are also easy to interpret and manipulate. Visual attributes have proven useful in face verification [7] and object classification [5], [8], particularly in the context of zero-shot learning [9], [6].

In many attribute-based problems [4], [10], [8], a (linear) transformation is learned so that *low-level* representations of images are projected into a high-level semantic space. Such a space is usually constructed so that each dimension describes the degree of presence of an attribute in an image. In other words, an image is described by a vector, and each element of the vector is the degree of presence of a given attribute in the image (see Section IV for details). In the attribute space, images can be semantically compared to one another. One of the most popular contexts that compare images with attributes is the relative attribute problem [4]. In this problem, the representations of images in the high-level semantic space are learned relatively to the learned representations of other images. The relative attribute problem considers relations between pairs of categories:

- inequality constraints: *i.e.* $(e) \prec_a (f)$: the presence of attribute a is stronger in category (f) than in category (e)
- equivalence constraints: *i.e.* $(g) \sim_a (h)$: the presence of attribute a is equivalent in category (g) and category (h).

This type of relationship is particularly useful when a Boolean score for the presence of an attribute is difficult to annotate. For instance, in Fig. 1, it may be difficult to annotate whether the image in category (e) is natural or not. Relative attributes tackle this problem by allowing people to annotate

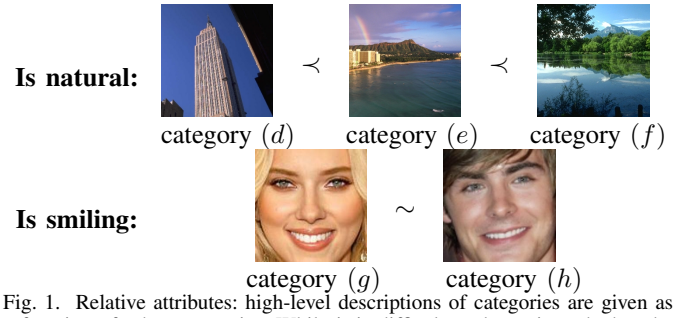

Fig. 1. Relative attributes: high-level descriptions of categories are given as a function of other categories. While it is difficult to determine whether the image in category (e) is natural, it is easier to say that it is more natural than category (d) and less natural than category (f). Scarlett Johansson (category (g)) smiles as much as Zac Efron (category (h)).

that the image in category (e) is more (or less) natural than the image in category (d) (or (f)).

B. Representations of Orders

In [4], the degree of presence of a given attribute is learned as a weighted sum, which induces an order over images with respect to that attribute. However, this aggregation function may not be the most appropriate representation for some attributes. In fact, in general, the choice of an aggregation function should depend on the attribute. In this section, we present different alternatives to the weighted sum model.

Mathematical representations of orderings have been theoretically investigated in many domains such as measurement theory [2], decision theory [11] or social sciences [12]. In this line of work, when the order $\prec$ is defined over objects that are represented by vectors in $\mathbb{R}^d$, an aggregation function $f : \mathbb{R}^d \to \mathbb{R}$ is said to *represent* $\prec$ if comparing objects via the values they receive from f leads to the same order as $\prec$. For instance, the weighted sum model is used as an aggregation function to represent relative attributes [4].

Many representations of orders[1] (and preorders[2]) have been proposed and studied theoretically. The focus in those mathematical approaches has been to characterize those representations by providing their *axiomatizations*, which are necessary and sufficient conditions (called *axioms*) on the order relation for the use of a representation. For instance, using a weighted sum to represent a given binary relation over vectors implies that the relation is a total[3] preorder that satisfies an independence axiom (*i.e.* no interaction between components of a vector) and a continuity axiom (*i.e.* small changes in a vector lead to small changes in the ordering).

In decision theory, when the objects to be compared are represented as vectors, as is the case with relative visual attributes, the problem is referred to as multi-objective (or multi-criteria) decision-making [3], [13]. In this area, the aggregation function f is called a *decision criterion* and many different criteria have been proposed and studied: Ordered Weighted Averaging (OWA) [14], Weighted OWA [15], (augmented weighted) Chebyshev norm [16]... Most of them may not be appropriate for our problem as they are specifically designed for multi-objective decision-making. One notable exception is the criterion defined as the Euclidean distance of a vector to an ideal point, which has been used to define compromise solution [13]. In the TOPSIS methodology [17], this criterion is combined with the distance to an anti-ideal point.

Interestingly, this criterion can also be found in social sciences [12]. In this setting, called *multi-dimensional unfolding*, individuals and objects are represented as points in a joint (multidimensional) space. Given a choice set, an individual will choose the object that is closest to him/her. In this context, the reference point is the individual in that joint space and also represents an ideal object.

The reference-point model has also been used in some machine learning work. [18] used it in an active learning setting for eliciting a preference order while trying to minimize the query complexity. In our case, the embedding problem assigns Euclidean coordinates only to the reference point without modifying the Euclidean low-level representations of images. Unlike classic embedding methods that do not extend to new samples, new examples can be easily added and compared to the reference-point.

Contributions: In this paper, we investigate the use of the Euclidean distance to a reference point as a representation for relative visual attributes. Although the weighted sum criterion and the Euclidean distance to an ideal point have been extensively used, to the best of our knowledge, not much work relate those two criteria. In this paper, we provide some theoretical results to clarify the relation between them. We also consider a third aggregation function defined as the Euclidean distance to an anti-ideal point.

We show that the order induced by the Euclidean distance to a reference point (*e.g.* ideal or anti-ideal points) can be seen as a relaxation of the order induced by a weighted sum, and we give a characterization of the weighted sum with respect to the other aggregation functions in the case of strict orders. Therefore, the Mahalanobis(-like) distance, which generalizes the Euclidean distance, can be seen as a further relaxation to the weighted sum. Finally, we illustrate how these aggregation functions based on the Mahalanobis distances can be exploited in the context of relative attributes.

III. Reference-Point-Based Model

We formally present the reference-point model and its use in the setting of relative attributes as a motivating example.

We assume that the ordering relations provided on categories (*e.g.* in Fig. 1) can be extended to the images of these categories. For a pair of images (described as feature vectors) $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}_j \in \mathbb{R}^d$, the notation $\mathbf{x}_i \precsim_a \mathbf{x}_j$ (resp. $\mathbf{x}_i \prec_a \mathbf{x}_j$ or $\mathbf{x}_i \sim_a \mathbf{x}_j$) means that the degree of presence of attribute a in image i is not greater than (resp. smaller than or equivalent to) the one in image j.

One of the goals of decision theory (and related domains) has been to investigate representations of binary relations. For

[1] Binary relation $\prec$ is a (strict) *order* if it is *irreflexive* (*i.e.* $\forall x$, not $x \prec x$) and *transitive* (*i.e.* $\forall (x, y, z)$, $x \prec y$ and $y \prec z \Rightarrow x \prec z$).

[2] A *preorder* $\precsim$ is *reflexive* (*i.e.* $\forall x$, $x \precsim x$) and transitive.

[3] Binary relation $\precsim$ is *total* if $\forall (x, y)$, $x \precsim y$ or $y \precsim x$.

a function $f : \mathbb{R}^d \to \mathbb{R}$, component-wisely non-decreasing (called *aggregation function* or *decision criterion*), we say that f *represents* a preorder $\precsim$ over $\mathbb{R}^d$ if:

$$\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d,\ \mathbf{x}_i \precsim \mathbf{x}_j \Leftrightarrow f(\mathbf{x}_i) \leq f(\mathbf{x}_j) \tag{1}$$

When relation $\precsim$ is a total preorder, as assumed in this paper, this can be equivalently written with its asymmetric part $\prec$:

$$\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d, \mathbf{x}_i \prec \mathbf{x}_j \Leftrightarrow f(\mathbf{x}_i) < f(\mathbf{x}_j) \tag{2}$$

In the work on relative visual attributes, the preorder $\precsim_a$ for an attribute a is represented by a weighted sum, *i.e.* $f_{\mathbf{w}_a}(\mathbf{x}) = \mathbf{w}_a^\top \mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^d$ is the feature vector of an image and $\mathbf{w}_a \in \mathbb{R}^d$ is the weight vector associated to attribute a. For a fixed weight vector $\mathbf{w} \in \mathbb{R}^d$, we denote $\precsim_{\mathbf{w}}$ (resp. $\prec_{\mathbf{w}}$) the preorder (resp. order) induced by the weighted sum defined by $\mathbf{w}$. Therefore, we have $\precsim_a = \precsim_{\mathbf{w}_a}$ and $\prec_a = \prec_{\mathbf{w}_a}$. A binary relation $\precsim$ (or $\prec$) for which there exists $\mathbf{w} \in \mathbb{R}^d$ such that $\precsim = \precsim_{\mathbf{w}}$ (or $\prec = \prec_{\mathbf{w}}$) is said to be *weight-based*.

As explained previously, the weighted sum is not always suitable to represent an order. We focus in this paper on two distance-based representations, in which objects are compared with respect to their distances to a fixed reference point. Before focusing on the Mahalanobis-like distance (see below for definition), we consider the Euclidean distance: $\mathsf{d}(\mathbf{x}_i, \mathbf{r}) := \|\mathbf{x}_i - \mathbf{r}\|_2 = \sqrt{\sum_{k=1}^d (x_{i,k} - r_k)^2}$ where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of an object and $\mathbf{r} \in \mathbb{R}^d$ represents a reference point. Note that any strictly increasing transformation of f in Eq. (1) yields a representation inducing the same (pre-)order. Therefore, from now on, we will use the squared distance $\|\mathbf{x}_i - \mathbf{r}\|_2^2$ for convenience' sake.

The first representation that we consider is defined by choosing the reference point $\mathbf{r}$ to be an ideal point[4] $\overline{\mathbf{r}}$ and is given by choosing $g_{\overline{\mathbf{r}}}(\mathbf{x}) = -\|\mathbf{x} - \overline{\mathbf{r}}\|_2^2$ as a criterion:

$$\mathbf{x}_i \precsim_{\overline{\mathbf{r}}} \mathbf{x}_j \Leftrightarrow g_{\overline{\mathbf{r}}}(\mathbf{x}_i) \leq g_{\overline{\mathbf{r}}}(\mathbf{x}_j) \tag{3}$$

Informally, an object is ranked higher if it is closer to the ideal point. A (pre-)order that admits this representation is said to be *ideal-focused*.

Symmetrically, the second representation sets the reference point $\mathbf{r}$ to an anti-ideal point $\underline{\mathbf{r}}$ and is defined by choosing $h_{\underline{\mathbf{r}}}(\mathbf{x}) = \|\mathbf{x} - \underline{\mathbf{r}}\|_2^2$ as a criterion:

$$\mathbf{x}_i \precsim_{\underline{\mathbf{r}}} \mathbf{x}_j \Leftrightarrow h_{\underline{\mathbf{r}}}(\mathbf{x}_i) \leq h_{\underline{\mathbf{r}}}(\mathbf{x}_j) \tag{4}$$

It states that an object is preferred when it is further from an anti-ideal point. A (pre-)order that admits this representation is said to be *anti-ideal-focused*.

Although less commonly used in practice, we introduce anti-ideal-focused relations because they will help us understand the relation between the weighted sum and the distance-based criteria. Note that the three representations are not equivalent in general. Indeed, an order described by one of those three criteria may not be representable by another one. The weighted sum model and the reference-point-based model may not be equivalent because their indifference curves (*i.e.* points considered equivalent) are different (*i.e.* hyperplan vs. hypersphere). The distance-based criteria may not be equivalent depending on the position of the reference point. For instance, the case where an ideal point is the centroid of a cloud of points may not find any representation with an anti-ideal point.

We now prove some theoretical results that shed some light on the relation between these three representations. Let $\mathcal{V} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be a set of objects endowed with a preorder $\precsim$. A binary relation can also be viewed as a set, *e.g.* $\precsim \subseteq \mathcal{V} \times \mathcal{V}$. We adopt this view to state our results for conciseness' sake. First, when a preorder is both ideal-focused and anti-ideal-focused, it is weight-based (see definitions above):

Proposition 1. *If* $\exists\ \overline{\mathbf{r}}, \underline{\mathbf{r}}$, $\precsim = \precsim_{\overline{\mathbf{r}}} \cap \precsim_{\underline{\mathbf{r}}}$, *then* $\exists\ \mathbf{w}$, $\precsim = \precsim_{\mathbf{w}}$.

Proof. Let $\precsim$ be ideal-focused w.r.t. $\overline{\mathbf{r}}$ and anti-ideal-focused w.r.t. $\underline{\mathbf{r}}$. It is easy to check that $\precsim = \precsim_{\mathbf{w}}$ with $\mathbf{w} = \overline{\mathbf{r}} - \underline{\mathbf{r}}$. □

Unfortunately, in general the converse is not always true, but it does hold for strict orders:

Proposition 2. *If* $\exists \mathbf{w}, \prec = \prec_{\mathbf{w}}$, *then* $\exists \overline{\mathbf{r}}, \underline{\mathbf{r}}$, $\prec = \prec_{\overline{\mathbf{r}}} \cap \prec_{\underline{\mathbf{r}}}$.

Proof. Without loss of generality, we assume $\mathbf{x}_1 \prec \mathbf{x}_2 \prec \ldots \prec \mathbf{x}_n$. By assumption, there exists $\mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{w}^\top \mathbf{x}_1 < \mathbf{w}^\top \mathbf{x}_2 < \ldots < \mathbf{w}^\top \mathbf{x}_n$. Let $\delta_1 = \min_{i>j} \mathbf{w}^\top (\mathbf{x}_i - \mathbf{x}_j) > 0$. Let $\delta_2 = \epsilon + \max_{i,j} |\mathbf{x}_i^\top \mathbf{x}_i - \mathbf{x}_j^\top \mathbf{x}_j|$ with $\epsilon > 0$. Let $\underline{\mathbf{r}} = -\delta_2/(2\delta_1)\mathbf{w}$ and $\overline{\mathbf{r}} = -\underline{\mathbf{r}}$. One can check that $\prec_{\underline{\mathbf{r}}} = \prec_{\overline{\mathbf{r}}} = \prec_{\mathbf{w}}$. □

In the general case, Prop. 2 can be extended as follows:

Proposition 3. *If* $\exists \mathbf{w}, \precsim = \precsim_{\mathbf{w}}$, *then* $\exists \overline{\mathbf{r}}, \underline{\mathbf{r}}$,

(i) $\prec = \prec_{\overline{\mathbf{r}}} \cap \prec_{\underline{\mathbf{r}}}$ *and*

(ii) $\sim = \big(\precsim_{\overline{\mathbf{r}}} \cap \succsim_{\underline{\mathbf{r}}} \big) \cup \big(\succsim_{\overline{\mathbf{r}}} \cap \precsim_{\underline{\mathbf{r}}} \big)$.

Proof. If $\prec = \emptyset$ (*i.e.* all points in $\mathcal{V}$ lie on the same hyperplane), choose $\overline{\mathbf{r}}$ and $\underline{\mathbf{r}}$ on different sides of this hyperplane and proportional to $\mathbf{w}$. Otherwise, by Proposition 2, there exists $\overline{\mathbf{r}}, \underline{\mathbf{r}}$ such that $\prec = \prec_{\overline{\mathbf{r}}} \cap \prec_{\underline{\mathbf{r}}}$. It is then easy to check that for $\mathbf{x}_i \sim \mathbf{x}_j$, we can have neither $\mathbf{x}_i \prec_{\overline{\mathbf{r}}} \mathbf{x}_j$ and $\mathbf{x}_i \prec_{\underline{\mathbf{r}}} \mathbf{x}_j$, nor $\mathbf{x}_i \succ_{\overline{\mathbf{r}}} \mathbf{x}_j$ and $\mathbf{x}_i \succ_{\underline{\mathbf{r}}} \mathbf{x}_j$, which yields (ii). □

This proposition states that a weight-based preorder can be seen as the combination of two preorders, one ideal-focused and one anti-ideal-focused, where two images are ranked in the order specified by those two preorders if they agree, otherwise, the two images should be considered equally ranked. Moreover, Propositions 2 and 3 imply that when ranking with a weighted sum, one has in fact implicitly chosen two reference points: one ideal point and one anti-ideal point. Furthermore, it suggests that using an ideal-focused or an anti-ideal-focused relation alone gives more flexibility than using a weight-based relation.

As a side note, Propositions 1 and 2 provide a simple characterization of weight-based strict orders:

[4] An ideal (resp. anti-ideal) point represents the most archetypal or typical (resp. atypical) point for an attribute. This definition is related to multidimensional unfolding, should not be confused to the notion of ideal/anti-ideal points in multiobjective optimization.

Corollary 1. $\exists \mathbf{w},\ \prec = \prec_{\mathbf{w}} \iff \exists \overline{\mathbf{r}}, \underline{\mathbf{r}},\ \prec = \prec_{\overline{\mathbf{r}}} \cap \prec_{\underline{\mathbf{r}}}$

These (pre)orders can be extended to the Mahalanobis(-like) distance widely used in distance metric learning [19] and defined for all $\mathbf{x}_i$ and $\mathbf{r}$ in $\mathbb{R}^d$ as:

$$\mathsf{d}_{\mathbf{M}}(\mathbf{x}_i, \mathbf{r}) := \sqrt{(\mathbf{x}_i - \mathbf{r})^\top \mathbf{M} (\mathbf{x}_i - \mathbf{r})} = \mathsf{d}(\mathbf{L}\mathbf{x}_i, \mathbf{L}\mathbf{r})$$

where $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$ is a $d \times d$ symmetric positive semidefinite matrix. The Mahalanobis(-like) distance $\mathsf{d}_{\mathbf{M}}$ generalizes the Euclidean distance (which corresponds to the special case where $\mathbf{M}$ or $\mathbf{L}$ is the identity matrix) and thus provides even more flexibility to the model, which may help better order the objects. Interestingly, such a general distance can encode some interactions between components of vectors thanks to the matrix $\mathbf{M}$, which is not possible with the weighted sum model. Representing a preorder with $\mathsf{d}_{\mathbf{M}}$ requires determining both the parameters $\mathbf{M} \succeq 0$ and $\mathbf{r} \in \mathbb{R}^d$.

In the context of relative attributes, it is probably justified for some attributes to learn and represent their orderings over images with weighted sums. However, in this paper, we argue that for other attributes, it may make more sense to learn a representation based on an ideal or anti-ideal point. This is for instance the case for the attribute "Masculine-Looking" (see Section V-A) where a representation based on an ideal point (representing a male stereotype) may make more sense because it may be difficult to define an anti-ideal point that would work for both males and females.

IV. Describing Images with Reference Points

We now present our learning algorithm that learns both an (anti-)ideal point and the Mahalanobis-like distance from a training dataset that describes pairwise comparisons.

a) Learning anti-ideal points: We focus on the case where we learn the anti-ideal point $\underline{\mathbf{r}}_a \in \mathbb{R}^d$ and the Mahalanobis distance $\mathsf{d}_{\mathbf{M}_a}$ for a given attribute a by exploiting the results of Section III; the ideal point case is similar.

We denote $\mathcal{A}_a$ the set of training pairs $\{(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^d \times \mathbb{R}^d : \mathbf{x}_i \prec_a \mathbf{x}_j\}$ where $(\mathbf{x}_i, \mathbf{x}_j)$ are vector representations of images. In the case of an anti-ideal point, we want to find a reference vector $\underline{\mathbf{r}}_a$ that is closer to $\mathbf{x}_i$ than to $\mathbf{x}_j$ for every pair $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{A}_a$. Formally, we want $\underline{\mathbf{r}}_a$ and $\mathsf{d}_{\mathbf{M}_a}$ to satisfy the maximum number of the following constraints:

$$\forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{A}_a, \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \underline{\mathbf{r}}_a) + 1 \leq \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_j, \underline{\mathbf{r}}_a) \quad (5)$$

$$\Leftrightarrow \ell\left(1 + \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \underline{\mathbf{r}}_a) - \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_j, \underline{\mathbf{r}}_a)\right) = 0 \quad (6)$$

where 1 is a safety margin and the convex loss function $\ell : \mathbb{R}^d \to \mathbb{R}^d$ can be written $\ell(x) = \max(0, x)$. Finding the optimal values of $\underline{\mathbf{r}}_a$ and $\mathsf{d}_{\mathbf{M}_a}$ is a NP-hard problem. Instead, we then optimize the following biconvex problem inspired by ranking SVM [1], [20]:

$$\min_{\substack{\underline{\mathbf{r}}_a \in \mathbb{R}^d \\ \mathbf{M}_a \succeq 0}} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{A}_a} \ell\left(1 + \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \underline{\mathbf{r}}_a) - \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_j, \underline{\mathbf{r}}_a)\right) + \mu \ \|\underline{\mathbf{r}}_a\|_2^2 + \nu \ \mathrm{tr}(\mathbf{M}_a) \quad (7)$$

where $\|\underline{\mathbf{r}}_a\|_2^2$ and $\mathrm{tr}(\mathbf{M}_a)$ are regularization terms, and $\mu \geq 0$ and $\nu \geq 0$ are regularization parameters.

For ease of exposition, we do not consider constraints induced by the equivalence relation $\sim_a$. Those equality constraints could be added in the optimization problem in a straightforward manner.

Since Eq. (7) is a biconvex problem, we solve it by alternating the optimization over $\underline{\mathbf{r}}_a$ and $\mathbf{M}_a$. We first initialize $\mathbf{M}_a$ as the identity matrix, optimize the problem over $\underline{\mathbf{r}}_a$ and alternate the optimization over each variable by fixing the other variable (see supplementary material for details).

b) Learning ideal points: To train an ideal point $\overline{\mathbf{r}}_a$, we replace in Eq. (7) the loss $\ell\left(1 + \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \underline{\mathbf{r}}_a) - \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_j, \underline{\mathbf{r}}_a)\right)$ by $\ell\left(1 + \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_j, \overline{\mathbf{r}}_a) - \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \overline{\mathbf{r}}_a)\right)$ and optimize over $\overline{\mathbf{r}}_a$ instead of $\underline{\mathbf{r}}_a$.

c) Exploiting reference points for classification: We now explain how we exploit ideal and anti-ideal points $\overline{\mathbf{r}}_a$ and $\underline{\mathbf{r}}_a$ for each attribute a to perform classification.

As explained in Section II-A, many attribute-based approaches learn a linear transformation so that low-level representations of images are projected into a high-level semantic space. In the case of relative attributes [4], a weighted sum is learned for each attribute $a \in \{1, \cdots, A\}$ where A is the number of attributes. More precisely, a vector $\mathbf{w}_a \in \mathbb{R}^d$ is learned so that the (maximum number of) following constraints are satisfied: $\forall i, j, \quad \mathbf{x}_i \prec_a \mathbf{x}_j \Rightarrow \mathbf{w}_a^\top \mathbf{x}_i < \mathbf{w}_a^\top \mathbf{x}_j$. Eventually, from the low-level representation $\mathbf{x}_i \in \mathbb{R}^d$ of an image i, a high-level representation $\mathbf{h}_i = (h_{i,1}, \cdots, h_{i,A}) \in \mathbb{R}^A$ is created where $h_{i,a} = \mathbf{w}_a^\top \mathbf{x}_i$. We propose different formulations of $h_{i,a}$ depending on some criteria that we explicit below.

Let $\mathcal{T}_a$ be a test set (different from the training set $\mathcal{A}_a$) composed of pairs $(\mathbf{x}_i, \mathbf{x}_j)$ such that $\mathbf{x}_i \prec_a \mathbf{x}_j$. Naturally, the best model for the attribute a between the weighted sum (WS), the ideal and the anti-ideal based representations (resp. IR and AR) is the one that best satisfies the following constraints over $\mathcal{T}_a$ (see Section V-A for details):

$$\begin{aligned} &\mathbf{w}_a^\top \mathbf{x}_i < \mathbf{w}_a^\top \mathbf{x}_j && \text{for WS} \\ &\mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \underline{\mathbf{r}}_a) < \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_j, \underline{\mathbf{r}}_a) && \text{for AR} \\ &\mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \overline{\mathbf{r}}_a) > \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_j, \overline{\mathbf{r}}_a) && \text{for IR} \end{aligned} \quad (8)$$

We then first consider that $\mathsf{d}_{\mathbf{M}_a} = \mathsf{d}$ and run 100 different train/test splits, the model that best satisfies these constraints for most splits is the model chosen for the attribute a since this means that it describes more accurately relations for the attribute. More precisely, for a low-level image representation $\mathbf{x}_i$ and a given attribute a, we formulate:

$$h_{i,a} = \begin{cases} \mathbf{w}_a^\top \mathbf{x}_i & \text{if WS is chosen} \\ \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \underline{\mathbf{r}}_a) & \text{if AR is chosen} \\ \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \overline{\mathbf{r}}_a) & \text{if IR is chosen} \end{cases} \quad (9)$$

Our high-level representation of an image i is then $\mathbf{h}_i = (h_{i,1}, \cdots, h_{i,A}) \in \mathbb{R}^A$ where $h_{i,a}$ is formulated as in Eq. (9). In the end, the high-level representation of the training data is used as the input of a classifier (*e.g.* linear SVM).

	Ideal Point		Anti-ideal Point	
OSR Attributes	Abs. diff.	Frob. norm	Abs. diff.	Frob. norm
Natural	24	46	22	37
Open	**67**	**73**	**60**	**65**
Perspective	6	12	29	41
Large-Objects	16	25	**85**	**90**
Diagonal-Plane	15	22	**51**	**62**
Close-Depth	**82**	**81**	**92**	**88**
PubFig Attributes	Abs. diff.	Frob. norm	Abs. diff.	Frob. norm
Masculine-Looking	**72**	**86**	43	**70**
White	45	**73**	**72**	**84**
Young	29	**56**	**80**	**86**
Smiling	12	41	**65**	**78**
Chubby	34	**65**	**63**	**77**
Visible-Forehead	35	**52**	19	38
Bushy-Eyebrows	46	**54**	27	**51**
Narrow-Eyes	45	**69**	16	28
Pointy-Nose	**53**	**69**	41	**61**
Big-Lips	23	**55**	**59**	**80**
Round-Face	28	**52**	**80**	**83**

TABLE I
COMPARISON OF THE REFERENCE POINT METHODS WITH THE WEIGHTED SUM ON THE TEST SET OVER 100 SPLITS. HIGHER IS BETTER.

V. EXPERIMENTAL RESULTS

To evaluate our reference point model, we follow a classification framework inspired by [4]. We experiment with the two datasets they used: Outdoor Scene Recognition (OSR) [21] containing 2688 images from 8 scene categories and a subset of Public Figure Face (PubFig) [7] containing 771 images from 8 face categories. We use the image features made publicly available by [4]: a 512-dimensional GIST [21] descriptor for OSR and a concatenation of the GIST descriptor and a 45-dimensional Lab color histogram for PubFig. Relative orderings of categories according to semantic attributes are given in [4, Table 1].

We denote category(u) the set of images in category u. By abuse of notation, we write category$(u) \prec_a$ category(v) if the presence of attribute a in category v is stronger than in category u, which implies $\mathbf{x} \prec_a \mathbf{y}$ for any image $\mathbf{x}$ of category u and any image $\mathbf{y}$ of category v.

A. Determining the best model for each attribute

Setup: We detail how we determine the model that is most appropriate for each attribute to apply the classification framework described in Section IV. We denote $c = 8$ the number of categories in both datasets. For each attribute a, we consider the ground truth order matrix $\mathbf{G}^a \in \{0,1\}^{c\times c}$ exploiting annotations provided in [4, Table 1]. $\mathbf{G}^a = (G^a_{uv})_{1\leq u\leq c, 1\leq v\leq c}$ is defined as:

$$G^a_{uv} = \begin{cases} 1, & \text{if category}(u) \prec_a \text{ category}(v) \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

For each model, we construct the matrix $\mathbf{F}^a \in [0,1]^{c\times c}$:

$$F^a_{uv} = \begin{cases} p^a_{uv} & \text{if category}(u) \prec_a \text{ category}(v) \\ p^a_{uv} = 1 - F^a_{vu} & \text{if category}(u) \succ_a \text{ category}(v) \\ 0 & \text{otherwise} \end{cases}$$

where $p^a_{uv} \in [0,1]$ is the ratio/fraction of pairs in the test set $\mathcal{T}_a$ that satisfy Eq. (8) for all $\mathbf{x}_i \in$ category(u), $\mathbf{x}_j \in$ category(v). We denote $\mathbf{F}^a_{WS}$ (resp. $\mathbf{F}^a_{ideal}$ or $\mathbf{F}^a_{anti-ideal}$) the matrix $\mathbf{F}^a$ obtained when using WS (resp. IR or AR).

Fig. 2. Different categories ordered *w.r.t.* the degree of presence of the attribute "smiling".

It is clear from the formulation above of $\mathbf{F}^a$ that the most appropriate model is the one for which the constructed matrix $\mathbf{F}^a$ is the closest to the ground truth matrix $\mathbf{G}^a$. To measure the discrepancy between $\mathbf{F}^a$ and $\mathbf{G}^a$, we use the following metrics:

1. absolute difference $\sum_{uv} |F^a_{uv} - G^a_{uv}|$
2. (squared) Frobenius norm $\|\mathbf{F}^a - \mathbf{G}^a\|^2_F = \sum_{uv}(F^a_{uv} - G^a_{uv})^2$

Table I reports for each attribute the number of times over 100 random training/test splits that $\mathbf{F}^a_{ideal}$ and $\mathbf{F}^a_{anti-ideal}$ are closer to $\mathbf{G}^a$ than $\mathbf{F}^a_{WS}$ is *w.r.t.* both metrics (*i.e.* if both methods have scores smaller than 50, then the weighted sum is the most appropriate for attribute a). Due to the small size of the datasets, we extract 80 and 50 training images per category on the OSR and Pubfig datasets, respectively. From these images, we create all the possible combinations of image pairs and use the annotations in [4, Table 1] to create our training constraints (based on $\mathcal{A}_a$ for reference point methods, see for example Eq. (5)) and we train the different models. The test set $\mathcal{T}_a$ is composed of all the possible pairs of the remaining images. When both the ideal and anti-ideal methods outperform the weighted sum, the reference point method with highest score actually outperforms the other one.

As can be seen in Table I, the anti-ideal point method outperforms the two other methods for half of the attributes *w.r.t.* both evaluation metrics on both datasets. On the other hand, the ideal point method outperforms the two other methods only for 2 attributes ("Masculine-Looking" and "Pointy-Nose") on Pubfig and 1 attribute ("open") on OSR.

Interpretation: We now explain for some attributes why the reference point models are more appropriate:

• *Smiling:* As illustrated in Fig. 2, although a relative ordering can be found between the different persons to rank the degree of presence of smile, there exist different kinds of smiles in the dataset. People smile with a closed mouth in the first row and with an open mouth in the second row of Fig. 2. While there are different kinds of smiles, the emotionless expression of Jared Leto (left) remains the same and corresponds to the anti-ideal point of smiling person. In this context, the anti-ideal point approach is more appropriate.

• *Masculine-Looking:* the Pubfig dataset is biased towards men since there are 6 categories of men and 2 categories of women. Among the 6 men, it is difficult to annotate whether some man is more masculine than some other man. Annotators then ranked the presence of "Masculine-Looking" according to their socially-defined stereotype of masculinity which is close to Clive Owen (third column of Fig. 2).

	Method	PubFig dataset	OSR dataset
	Weighted sum	$75.0 \pm 0.4\%$	$69.6 \pm 0.4\%$
$\mathsf{d}_{\mathbf{M}_a}=\mathsf{d}$	Ideal point (with d)	$74.7 \pm 0.6\%$	$69.2 \pm 0.5\%$
	Anti-ideal point (with d)	$77.9 \pm 0.7\%$	$72.0 \pm 0.9\%$
	Combination (with d)	$\mathbf{78.7 \pm 0.4\%}$	$\mathbf{72.8 \pm 0.7\%}$
$\mathsf{d}_{\mathbf{M}_a}\neq\mathsf{d}$	Ideal point (with $\mathsf{d}_{\mathbf{M}_a}$)	$75.7 \pm 1.2\%$	$\mathbf{73.8 \pm 0.6\%}$
	Anti-ideal point (with $\mathsf{d}_{\mathbf{M}_a}$)	$78.0 \pm 1.0\%$	$72.1 \pm 0.7\%$
	Combination (with $\mathsf{d}_{\mathbf{M}_a}$)	$\mathbf{82.1 \pm 1.0\%}$	$73.2 \pm 0.6\%$

TABLE II
TEST CLASSIFICATION ACCURACY (MEAN AND STANDARD ERROR) FOR THE DIFFERENT HIGH-LEVEL IMAGE REPRESENTATIONS.

Similar explanations can be given for other attributes. For instance, only one person has a pointy nose in the PubFig dataset and the ideal point model is more appropriate because of the presence of different kinds of non-pointy noses (*e.g.* flat or round) in the dataset.

B. Classification results

We compare the classification performances of the different reference point strategies to construct high-level image representations that are used as input of a linear SVM classifier. We use the same number of training images as in the previous task and run the experiments on 10 new random training/test splits. We report in Table II the average classification accuracy across categories (*i.e.* mean of the accuracies obtained for each category).

We consider the weighted sum baseline which corresponds to the method proposed in [4]. It considers the high-level representation $\mathbf{h}_i = (h_{i,1}, \cdots, h_{i,A}) \in \mathbb{R}^A$ where $h_{i,a} = \mathbf{w}_a^\top \mathbf{x}_i$.

Additionally to this baseline, we compare in Table II three reference-point-based strategies to construct $\mathbf{h}_i \in \mathbb{R}^A$:
1. The ideal point method considers $\forall a, h_{i,a} = \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \overline{\mathbf{r}}_a)$ where $\overline{\mathbf{r}}_a$ is the learned ideal point for attribute a.
2. The anti-ideal point method corresponds to $\forall a, h_{i,a} = \mathsf{d}^2_{\mathbf{M}_a}(\mathbf{x}_i, \underline{\mathbf{r}}_a)$ where $\underline{\mathbf{r}}_a$ is the learned anti-ideal point.
3. The combination of the best representations for each attribute as described in Eq. (9). The best method is determined by exploiting the scores in Table I: if both the ideal and anti-ideal point methods have at least one score smaller than 50 in Table I, then the weighted sum is chosen.

We report the scores for cases where the Mahalanobis distance is learned with reference points (*i.e.* $\mathsf{d}_{\mathbf{M}_a}\neq\mathsf{d}$) and where it is not learned (*i.e.* $\mathsf{d}_{\mathbf{M}_a}=\mathsf{d}$). The ideal point method, which already obtains the worst scores in Table I, usually achieves the worst performance in classification accuracy. However, the anti-ideal point method which obtains the best scores in Table I for half of the attributes outperforms the weighted sum. This demonstrates a correlation between the scores in Table I and suggests that better representations lead to more accurate classification. The appropriate combination of the best representations for each attribute further improves classification performance, which validates our reference point approach. Learning a metric with the reference point slightly improves results.

VI. CONCLUSION

We have proposed and justified the use of reference-point-based decision models, which can be seen as a relaxation of the classic weighted-sum model, to deal with ordered relations between objects or categories. Particularly, we have successfully applied our reference-point-based approach in the context of relative visual attributes where our method seems more appropriate than the weighted sum for some attributes. Extensions to other types of attributes (*e.g.* related to ethnicity) such as relative similes [7] (*e.g.* similarity with Halle Berry's nose or Robert Redford's mouth) are straightforward. Furthermore, although we illustrate our approach for the relative attribute problem, it could be applied to other contexts such as late fusion in multimodal problems [22].

REFERENCES

[1] T. Joachims, "Optimizing search engines using clickthrough data," in *SIGKDD*. ACM, 2002, pp. 133–142.
[2] D. Krantz, D. Luce, P. Suppes, and A. Tversky, *Foundations of measurement*. Academic Press, 1971, vol. Additive and Polynomial Representations.
[3] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, 1976.
[4] D. Parikh and K. Grauman, "Relative attributes," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.
[5] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*. IEEE, 2009.
[6] M. Bucher, S. Herbin, and F. Jurie, "Improving semantic embedding consistency by metric learning for zero-shot classification," in *European Conference on Computer Vision (ECCV)*, 2016.
[7] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 365–372.
[8] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 819–826.
[9] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *NIPS*, 2009.
[10] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *CVPR*. IEEE, 2013, pp. 771–778.
[11] S. Barberà, P. Hammond, and C. Seidl, *Handbook of Utility Theory*. Springer, 1999.
[12] C. H. Coombs, *A theory of data*. Wiley, 1964.
[13] P.-L. Yu, *Multiple-criteria decision making: concepts, techniques, and extensions*. Springer Science & Business Media, 2013, vol. 30.
[14] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 18, no. 1, pp. 183–190, 1988.
[15] V. Torra, "The weighted owa operator," *International Journal of Intelligent Systems*, vol. 12, no. 2, pp. 153–166, 1997.
[16] R. E. Steuer and E.-U. Choo, "An interactive weighted tchebycheff procedure for multiple objective programming," *Mathematical programming*, vol. 26, no. 3, pp. 326–344, 1983.
[17] C. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications*. Springer-Verlag, 1981.
[18] K. G. Jamieson and R. Nowak, "Active ranking using pairwise comparisons," in *NIPS*, 2011.
[19] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *NIPS*, 2002, pp. 505–512.
[20] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Information Retrieval*, vol. 13, no. 3, pp. 201–215, 2010.
[21] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision (IJCV)*, vol. 42, no. 3, pp. 145–175, 2001.
[22] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *ACM Multimedia*. ACM, 2005.