ORIGINAL ARTICLE

A semi-automatic mold cost estimation framework based upon geometry similarity

Cyrus Hillsman · Yan Wang · Dima Nazzal

Abstract The automation of cost estimation for manufacturing processes is a challenging task in computer-aided manufacturing. In this paper, we introduce a two-step analogy and mathematical approach to estimate the cost of injection molding. In the analogy step, data of molds are partitioned into homogeneous groups based on mold type and mold design. In the prediction step, regression models based upon geometry, topology, and other inherent shape properties are constructed within each group. The variables in the regression models within each group are extracted automatically from one orthographic two-dimensional (2D) image of the injection-molded part. Mean and variance estimates are calculated on a subset of relevant molds so that the risk of an inaccurate bid can be assessed on a subset of relevant molds.

Keywords Injection molding · Cost estimation · Geometry similarity · Wavelets · Image processing · Topological descriptors · Pattern recognition

1 Introduction

The North American Industry Classification System (NAICS) code 3261, "Plastics Product Manufacturing",

C. Hillsman (✉)
Veterans Administration Center for Applied Systems Engineering,
2669 Cold Spring Road,
Indianapolis, IN 46222, USA
e-mail: cyrushillsman@hotmail.com

Y. Wang
George W. Woodruff School of Mechanical Engineering,
Georgia Institute of Technology, Atlanta, GA 30332, USA

D. Nazzal
H. Milton Stewart School of Industrial & Systems Engineering,
Georgia Institute of Technology, Atlanta, GA 30332, USA

accounted for 163 billion dollars in 2005. Throughout the world, the competitive bid process is of particular interest since many injection-molding firms still operate as job shops. The reason for this interest is that the mold and related design comprise 50 % of the total cost of a typical injection-molded part over its lifetime. Significantly, the industry still lacks a universally accepted method to bid the injection molds. The most common methods used to estimate the costs of injection molds are ad hoc and heavily rely on personal experiences and subjective judgments.

Some research has been undertaken to estimate costs in a more systematic fashion. But in general, they have not sought to automate the process and have not considered the variability of cost information. Most of the research efforts are analogy-based; therefore, the estimations tend to be qualitative. In addition, significant cost-contributing factors, e.g., the complex internal features of some mold and part designs, have been neglected which affects the accuracy of the predictions.

In this paper, we propose a new hybrid approach that combines analogy-based clustering and regression models. Similar molds are clustered into groups first. Then regression models are built to estimate cost based on mold complexity. The complexity is quantified by shape descriptors derived from two-dimensional (2D) images of injection-molded parts. Particularly, wavelet descriptors of boundaries as well as other inherent shape descriptors, including size and number of boundaries, are used to describe the complexity of each part. The descriptors are then used to build regression models. The objective is the creation of a cost-effective and standardized methodology for automatic or semi-automatic cost estimation of injection molds considering part complexity and variation.

Our hybrid approach is unique for three reasons. First, the previous methods only consider one data format and fail to make use of all of the historical data available. Our approach

uses images that can be created from traditional blueprints, 2D, or 3D computer-aided design (CAD) data. Second, the existing research only partially classify the mold and part data into groups of mold types, designs, and part complexities qualitatively prior to cost estimation. In contrast, our approach incorporates internal features as an important component of geometry similarity measurement so that the cost estimation is based on fine-grained quantitative classifications. Third, the previous methods use point estimates and do not evaluate the variations of mold costs thereby neglecting risk management, which is considered as an important element in our approach.

In the remainder of the paper, Section 2 provides a brief introduction of mold cost estimation, geometry similarity, and wavelets. In Section 3, our proposed cost estimation framework is presented. In Section 4, the implementation and experimental results are described.

2 Background

In the injection molding process, solid plastic pellets approximately the size of a grain of rice are heated until they liquefy and then injected into the mold under pressures of 5,000–15,000 psi. After the plastic is injected into the mold, the plastic cools to a solid state and is ejected from the mold in preparation for the next cycle. There are three main types of injection molds, each with a different cost structure. The first is a *conventional* mold where one mold is dedicated to one part. The second type is *master unit dies* (*MUD*), which requires less mold base preparation. The third type requiring the least preparation is a *modular* mold, where each part has its own insert in the master mold base. For the three mold types, three mold designs are commonly used. The least expensive design type is *straight*, whereby an ejector pin driven by the action of the injection-molding machine ejects the part from the mold. The second mold design type is

spring-loaded ejector, whereby the ejector system is self-contained within the mold. The third design type is *cam action*, whereby the demolding requires cams and springs to release the part. Cam action is the most expensive mold design.

2.1 Mold cost estimation methods

The most commonly used approach for estimating injection molds in industry is ad hoc and is solely based on the experiences of the bidder who prepares the bid. Academic researchers have tried to develop more systematic methods to evaluate the cost of molds. Table 1 provides an overview of the existing research used to estimate the cost of injection molds. The first category is analogical, whereby molds are first grouped and the mold to be bid is compared to those ones within the same group. The rationale is that similar molds cause similar costs. The second category is prediction by mathematical models with parameters such as the number of features of the part and the number of cavities. Both methods have advantages, and our method seeks to bridge the gap between the two methods. Our hybrid approach incorporates both analogy and mathematical models. The analogy-based method that is closest to our approach is by El-Mehalawi [1]. But it differs because constructive solid geometry was used as the input data, the descriptor for clustering was based on the attributed graph, and the method was not validated.

In our clustering process, we first group the jobs together by mold type (conventional, MUD, or modular). Next, the jobs in each group are further clustered based upon the design of the mold (straight, spring ejector, or cam action). Each group is categorized even further according to geometry and complexity, which provides a finer granularity. After clustering, the dataset becomes sparse and reflects only those most similar molds for estimation thereby providing for more accurate estimation of mean and variance on relevant molds.

Table 1 Comparison of cost estimation research projects

Author	Type	Automated part matching	Validated	Implemented
El-Mehalawi [1]	Analogy	Yes	No	Yes
Kwong and Smith [2]	Analogy	No	No	Yes
Wong [3]	Analogy	No	No	Yes
Wang et al. [4]	Analogy	No	No	Yes
Bergman [5]	Analogy	No	Yes	Yes
Shehab and Abdalla [6]	Mathematical	No	No	Yes
Nagahanumaiah et al. [7]	Mathematical	No	Yes	Yes
Raviwongse and Allada [8]	Mathematical	No	No	Yes
Shehab and Abdalla [9]	Mathematical	No	No	Yes
Chan et al. [10]	Mathematical	No	No	Yes
Sapene [11]	Mathematical	No	No	No

2.2 Geometry similarity

Geometry similarity provides the finest detail of cost estimation. There are several active research groups working in the area of geometry similarity representative papers from the research groups are as follows: Temple Shape Similarity Project [12], Princeton Shape Retrieval and Analysis Group [13], and Purdue PRECISE lab [14]. The major research question is how to quantify geometry similarity and cluster shapes. The research methods can be divided into two categories, 2D shape similarity and 3D shape similarity. 2D shape similarity is well studied and represented in the literature. Example metrics to quantify similarity are shading, boundaries of silhouette, distributions of distance fields, statistical moments, and Fourier descriptors. Reviews can be found in Belongie et al. [15] and Loncaric [16]. The interest in 3D similarity has grown significantly in recent years. Example metrics are based on polyhedral meshes and 3D Fourier descriptors as the extension of 2D methods. Comprehensive surveys for 3D shape similarity are given by Iyer et al. [17], Tangelder et al. [13], Cardone et al. [18], and Bustos et al. [19].

In our work, wavelet descriptors are used to represent 2D boundary information of geometry that is extracted from 2D images, with a different goal than the above approaches. Similar parts are grouped based upon the wavelet descriptors. Wavelets analysis offers the advantage of multi-resolution, which is introduced in the next section.

2.3 Wavelets and wavelet descriptors

A good introduction to wavelets can be found in Walker [20]. Wavelets are the basis functions used in the wavelet transform, which decomposes a vector space with multiple resolutions. The difference between the complete vector space V_0 and the scaling function subspace V_1 is defined as the wavelet space W_1. Wavelets and the associated scaling functions constitute a basis for the complete vector space where $V_0 = V_1 \oplus W_1$. The wavelet and scaling functions are orthogonal complements and form a complete basis, that is, $\forall w \in W_1, \forall v \in V_1, w|v = 0$. The scaling function subspaces form a multi-resolution analysis of the complete vector space, i.e., $V_0 \supset V_1 \supset V_2 \supset \cdots \supset V_K$ for $K+1$ levels of resolutions. At each level, the scaling function and wavelet can be used to construct the next level. That is, $V_i = V_{i+1} \oplus W_{i+1}$ for $i = 0, \ldots, K-1$.

Wavelet descriptors used in our method are based upon the wavelet function coefficients h_ψ and scaling function coefficients h_φ. They offer three major advantages over other shape descriptor methods. First, the inherent multi-resolution representation makes shape comparison efficient since our goal is to build a parsimonious model using the least possible number of coefficients. Second, they allow data compression which can help speed up the process of clustering. Third, they are more robust than those based on global coefficients, such as the Fourier transform. Local changes to a signal remain local after the wavelet transform. A local feature only affects those coefficients that are local to that region.

3 The proposed cost estimation framework

The overall framework of our proposed method is illustrated in Fig. 1. First, all part data are converted to a common format of 2D images. Then the mold type and the mold design are selected and the part complexity is calculated. The dataset of all molds are clustered into subsets of relevant molds so that only those most similar mold designs are used for cost estimation. For each subset of data, three types of descriptors (regional, topological, and wavelet) are combined into one feature vector for each part along with the mold type and design data. Lastly, the combined feature vector is used as the basis to construct a regression model for prediction. Each of these steps is explained in the following sections.

3.1 Conversion of data to a common neutral format

One unique feature of our approach is that the estimations of the mold costs are calculated directly from part drawings, where all part design data are converted to a neutral format of images prior to further processing. There are several reasons to choose 2D image as the neutral format. First, 2D images can be easily accessed and visualized without the use of complex software tools, which are not necessarily available in small and medium enterprises. Second, 2D images provide ease of conversion from other data formats. Third, the data processing power required to display images is less than CAD software. Image processing tools are much more matured and readily available as commodities. Fourth, 3D representation is not necessarily more powerful than 2D representation in capturing geometry information for the purpose of feature recognition [21].

3.2 Mold type selections

The selection of mold type in our approach is manual and is based upon both geometric and non-geometric information. The non-geometric information such as estimated annual usage, material, and application is as important as geometry to the selection process. Therefore, this selection process requires the participation of an experienced bidder who is familiar with the mold types. Here, the selection of mold type is restricted to conventional, MUD, and modular because they are the most common types and encompass a large proportion of all molds used in industry.
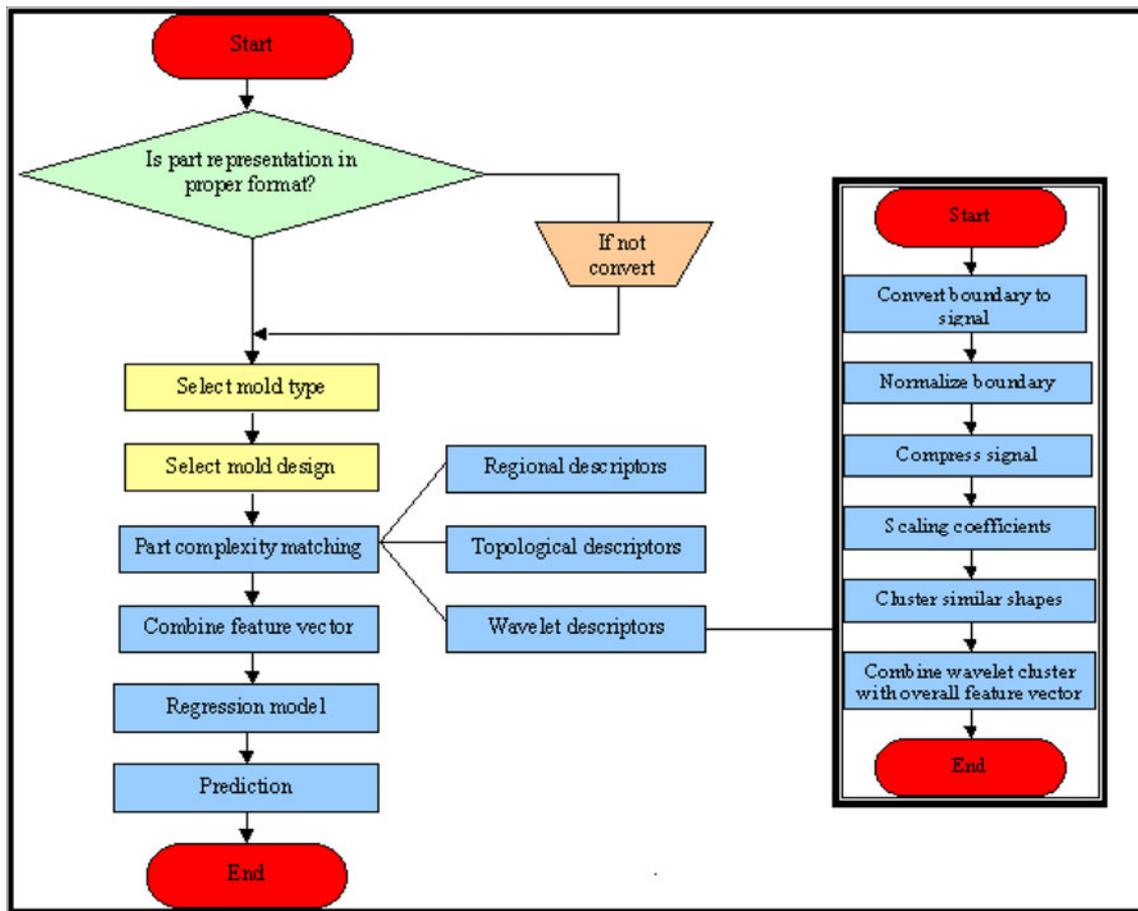
Fig. 1 The process of cost estimation in the proposed framework

3.3 Mold design selections

Mold design selection process also heavily relies on the experts' experiences today. Over the past decades, the knowledge of mold design and construction has been accumulated and books are available such as the one by Kluz [22]. During design, part demolding and part ejections are the primary considerations. It is believed that a manual selection process as in our method is still the most reliable approach. The full automation of this process does not necessarily bring benefits. Here mold designs are restricted to the three most commonly used ones: straight, spring ejector, and cam action.

3.4 Part complexity matching

The most important procedure of the proposed cost estimation framework is part matching. The complexity of a part is represented using several descriptors that are obtained from an image of the part. The process of part matching consists of three steps. First, the image is pre-processed and the boundaries of the part are detected. Here, we only use one view of the part, for instance,

the top view looking directly into the parting line. Second, the image is normalized by necessary transformations (translation, rotation, reflection, and scaling) for ease of comparison. Finally, each detected boundary is characterized by some shape descriptors.

3.4.1 Image pre-processing and boundary detection

The image pre-processing such as contrast adjustment and boundary detection can be done easily with some standard image-processing libraries such as the one in Matlab. A typical image after the pre-processing is shown in Fig. 2. The shaded area is one feature on the injection molded part. The feature is enclosed by a boundary represented as a circle. Boundaries are those connected components, which define features on the injection molded part. After the images are processed, the boundaries can be traced. Then segmentation is performed so that the 2D image is divided into regions as described in Gonzalez [23].

In an example shown in Fig. 3, the part has three distinct boundaries that enclose regions indicated by different shades and colors. The first boundary is the gear-shaped
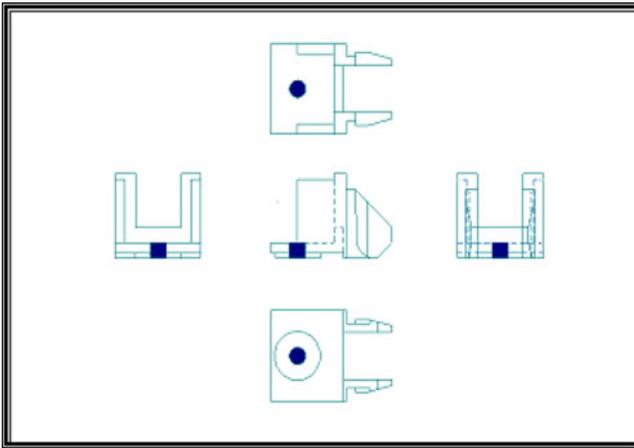
Fig. 2 A typical image of part

outer one. The second boundary is internal and has a reversed "C" shape. The third one is also internal and has a circular shape with a tab on the left side. The number of boundaries is related to the complexity of the part. Each boundary is considered to be associated with a feature of the part.

3.4.2 Normalization

The normalization process first translates the image to the origin with zero x and y coordinates and crops the image. Then the image is rotated so that the major axis of the part aligns with the x-axis. Mirror or reflection compares the center of the bounding box to the center of mass of the boundary. Finally, the scale of the image is normalized so that a common pixel-to-length ratio is maintained, for example, 1,000 pixels per 1 in. of length.
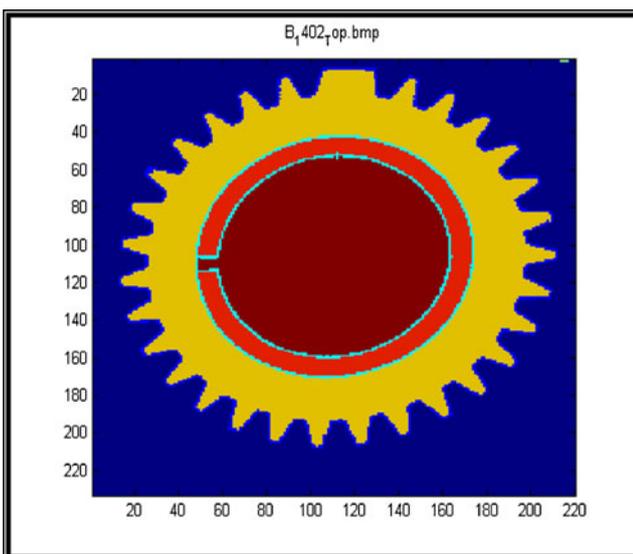

Fig. 3 An example of boundaries and regions

3.4.3 Shape descriptors of boundaries

It is important to note that the use of geometry is an indirect way to measure complexity of molds and complexity is correlated with cost. Therefore, our goal is slightly different from the pure shape similarity assessment as mentioned in Section 2.2. We want to create a feature vector that describes the complexity of all boundaries resulted from the previous steps so that clustering of geometry with respect to cost can be achieved. Each boundary is characterized by a set of descriptors.

There are three categories of descriptors to measure shape complexity. The first category includes the regional descriptors such as size, eccentricity, bounding box, etc. The second one includes topological descriptors, particularly the number of enclosed boundaries, Euler numbers, Betti Numbers, and Genus. The third one is through coefficients of wavelet transform. They are introduced respectively as follows.

1. Regional descriptors

   Regional descriptors are the local properties that describe a boundary's basic geometric information. For instance, the area of a region in a 2D image is related to the volume of the feature therefore the amount of materials added or subtracted during fabrication. The eccentricity is a metric of how similar the boundary is to a circle, which is calculated by the ratio of the distance between the foci of the ellipse and its major axis. The size of the bounding box for the region is also related the volume of the feature.

   We chose six regional descriptors, including *area*, *eccentricity*, *convex area*, *filled area*, *extent*, and *solidity*. The primary reason we selected them is that they can be represented as scalar values and easily combined. *Area* is the number of pixels in a region. *Eccentricity* is the ratio of the distance between the foci of the ellipse and its major axis. *Convex area* is the size of a convex polygon approximating the boundary shape. *Filled area* is the size of the bounding box. *Extent* is the area of the boundary divided by the area of the bounding box for that region. *Solidity* is the area divided by the convex area.

2. Topological descriptors

   Besides geometry, the second type of metrics to capture complexity is through topographical descriptors, including the Euler number and the number of boundaries in the image. The Euler number is the number of boundaries in the image minus the number of holes in those regions. The number of boundaries is an indication of how many features are with a part. Fewer boundaries would imply a simpler geometry of the part.

3. Wavelet descriptors

The third descriptor type is the wavelet descriptor. As shown in Fig. 1, extracting the wavelet descriptor consists of several steps. At this stage, we assume that the images have been pre-processed and the boundaries have been detected and normalized. A starting point of a boundary is identified first. The starting point is the pixel on the boundary with the shortest Euclidean distance to the origin after translation, rotation, and reflection.

The general procedure of wavelet descriptor construction is as follows: (1) choose which wavelet to use, (2) convert the 2D boundary to a 1D signal, (3) normalize the length of the signal, (4) compress the signal and calculate the scaling coefficients, and (5) cluster similar shapes together and make the cluster part of the overall feature vector for cost estimation.

For step 1 above, we use the Daubechies wavelet with four coefficients since the Daubechies wavelets have property II [20], which states that if J is the length of the wavelet filter and the signal is approximately a polynomial of degree less than $J/2$ over the support, then the wavelet coefficients will be approximately zero. In other words, the scaling function for the Daubechies wavelet with four coefficients approximates straight lines well. Therefore, a lower resolution subspace with fewer coefficients can still represent the shape accurately. This enables efficient compression.

In step 2, a boundary in a 2D image is converted to two 1D signals. The signal is a function of parameter t, which is the distance along the x-axis or y-axis to the starting point. Two simple examples are shown in Fig. 4. The signal $X(t)$ in Fig. 4b is the distance along the x-axis to the starting point (the bottom left corner) as the boundary is traced counterclockwise in Fig. 4a. Figure 4d is the x-axis distance of the circular boundary in Fig. 4c. Similarly, we can construct the y-axis distance signal $Y(t)$.

In step 3, the signal length is normalized. The signal is scaled so that it has a standard length, which is a power of two because the downsampling process in the wavelet transform reduces the length of signals by half.

In step 4, we compress the signal because it will reduce the time required for further processing. Even after compression, the signals retain their essential shape, which is captured by the scaling function coefficients. These coefficients contain the bulk of the energy of the signal and therefore the basic geometric information of the original signal.

In the last step of wavelet descriptor construction, the similar compressed signals are clustered. As a result, boundaries with similar shapes are grouped. The indices of wavelet clusters are used to construct the overall feature vector.

3.5 Create overall feature vector

Having collected the information on mold type, mold design, and part complexity, our final task is to create a comprehensive feature vector. It is desirable to build a parsimonious model that accurately reflects the relationship between the independent variables and cost.

The known cost estimate will serve as our target in the supervised learning method or regression. The mold feature vector is combined with part feature vectors for the descriptors of geometry and topology. The mold feature vector includes mold type and design. The part feature vector is formed by the topological, regional, and wavelet descriptors. All descriptors of the mold and part form one observation.

3.6 Regression models

Having created the feature vector, we can now build regression models. The multiple regression technique is chosen as the supervised learning method because it provides a way to combine multiple data types. The feature vector has a mixture of data types of continuous, ordinal, and categorical data. We want to discover how the variables of the selected descriptors are related to the dependent variable of cost. In addition to the prediction of mean value, variation prediction will help us determine how precise the estimate can be.
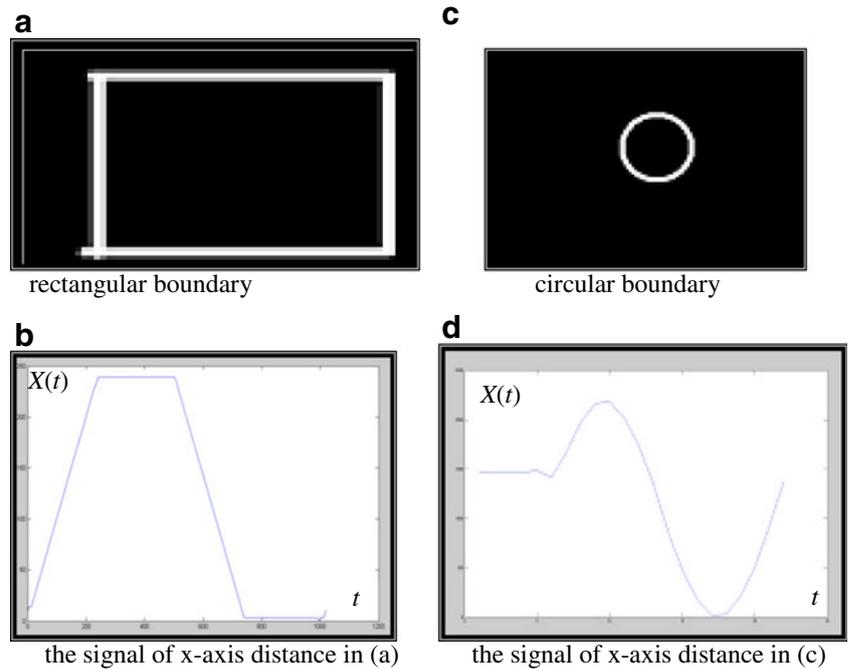
4 Implementation and experimental results

4.1 Implementation

In the implementation, the Microsoft Access database was used to store information about the mold and cost. Matlab was used to build the image processing portions of our method. Ultimately, all the information is combined into Minitab for the statistical and regression analysis.

Figure 5 shows the typical data automatically collected from the part print. *Name* is the unique name of the mold at the company which supplies the data. Columns 1 and 2 in the dataset named *record* are the total number of boundaries and the boundary indices for one image, respectively. The dataset labeled *T* is the cluster index of the wavelet descriptors for the boundaries. *T2* is the cluster index based upon the regional descriptors. In addition, *Eccent* is the eccentricity of the boundary as the additional variable. Through our investigation, we find that boundaries with eccentricity below 0.40 highly correlate with circular boundaries. Therefore, the variable *Eccent* is used to determine the number of non-circular boundaries (*NonCir*) in the image. Round boundaries are easily machined and have a lower cost than non-circular ones. Our implemented Matlab code extracts the topological, regional, and wavelet descriptors automatically.

Fig. 4 Boundaries converted to signals

a

rectangular boundary

c

circular boundary

b

$X(t)$

t

the signal of x-axis distance in (a)

d

$X(t)$

t

the signal of x-axis distance in (c)

4.2 Results

The dataset used in this study consists of 83 molds: conventional, master unit die, and modular type, which are all real-world examples collected from industry. There are 16 conventional molds, 18 master unit dies, and 49 modular molds. All mold and part prints are scanned to images, and images are processed so that boundary information can be detected. The result of the descriptor extraction process is one master dataset as shown in Fig. 6, which is used to build the regression models.

The master dataset used three sources of data: the invoices of the molds that include cost, the print of the mold, and the prints of the plastic parts themselves. The mold print provides the variable information about the number of cavities and the design of the mold. These variables are manually entered into the database. As described in Section 3, wavelet descriptors, regional descriptors, and topological descriptors are calculated automatically from the part print.

4.2.1 Mold-related variables

Four mold-related variables, *Name*, *NumCav*, *Type*, and *Design*, are taken from the mold print manually and entered into the database. *Name* is the name of the mold and is the unique identifier for that mold within the company that supplied the data. *Num_Cav* is the number of cavities in

Fig. 5 Matlab database
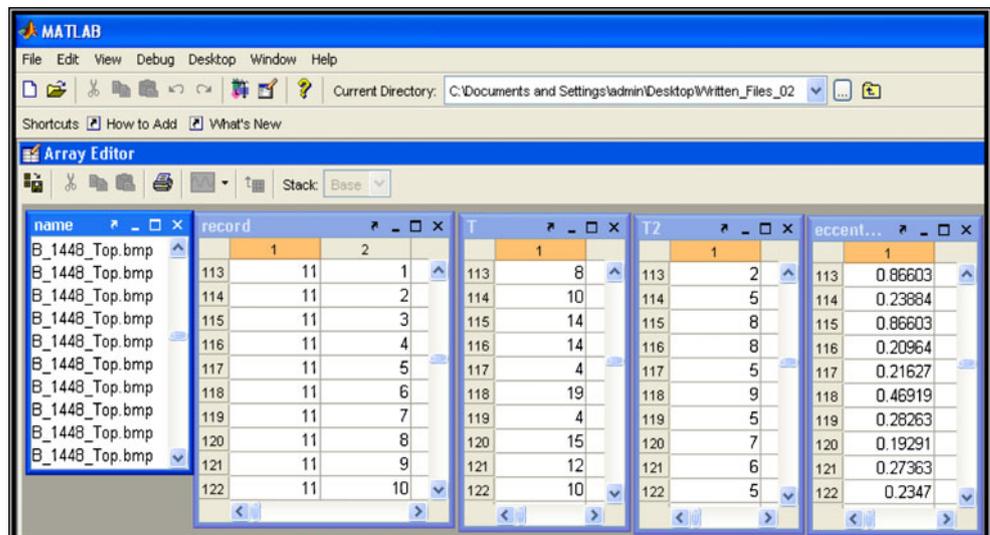
the mold. Intuitively, the more cavities a mold contains, the more labor and materials it requires. Although more labor and materials should result in higher cost, this variable is not in isolation. *Type* is the mold type taken from the mold print and represents the system used to manufacture the mold either conventional (*cov*), master unit die (*mud*), or modular (*mod*). As mentioned in Section 3.2, these are fundamentally different systems, each with a different cost structure. *Design* is the mold design named *straight*, *spring*, or *cam*. As described in Section 3.2, and 3.3, different mold types and engineering designs should serve as a partition for the dataset.

The dependent variable *Cost* is taken from the customer invoice, which is the quoted and final sale price. The prices are calculated based upon experience of the bidder. However, they may not reflect reality in all cases. Factors such as missed bids, intentional over- or under-bid, surface finish, or tolerance may also be included in the bid.

4.2.2 Part-related variables

As shown in Figs. 6 and 7, the scanned images of parts are processed, and the boundaries and regions are detected automatically, as indicated by the different colors. Four part-related variables, *NumBound*, *WaveClust*, *RegClust*, and *NonCir*, are computed from the part images semi-automatically. The four variables are explained in detail as follows.

NumBound is the number of boundaries in the image for a part. Intuitively, the more features a part has, the higher the manufacturing cost could be.

WaveClust is the total number of unique clusters within one part out of a total number of general shape clusters based on the wavelet scaling coefficients. Here, we define 20 shape categories for the given 83 parts and cluster all boundaries for all parts into the 20 clusters. Then we count the unique boundary clusters that a given part has as *WaveClust*.

Counting the unique clusters is an attempt to measure symmetry, as more symmetric parts are assumed to have less complexity for a given number of boundaries. As an example of a symmetric part, the part in Fig. 8 has 27 boundaries. However, some of these boundaries are in common clusters and have a similar shape. Therefore, the part has *WaveClust* = 3 unique boundary shapes. In contrast, a typical non-symmetric part is shown in Fig. 9.

The values for each individual regional descriptor are calculated for each boundary within an image. The six descriptors listed in Section 3.4.3 as well as the Euler number are clustered into ten approximate shapes for all boundaries for all parts in the dataset. We then record the cluster number for each boundary. *RegClust* is defined as the total number of unique clusters for each part out of the possible ten clusters.

Similar to *WaveClust*, *RegClust* is also an attempt to measure symmetry. The rationale is that if two boundaries

Fig. 6 Master dataset

|	Name	Num_Cav	Design	Type	Cost	Num_Bound	WaveClust	RegClust	NonCir
1	B-450	1	Cam	Mod	5500	14	9	7	9
2	B-454	1	Cam	Mod	7500	6	5	4	1
3	F-109	1	Cam	Mod	5325	11	8	5	4
4	F-114	1	Cam	Mod	8900	24	19	11	14
5	K-1457	1	Spring	Mod	2400	4	4	1	2
6	K-1474	2	Spring	Mod	3000	4	4	1	4
7	K-1481	1	Spring	Mod	4000	8	8	2	2
8	K-1482	1	Spring	Mod	4000	6	6	2	0
9	K-1483	1	Spring	Mod	4000	6	6	2	0
10	K-1487	1	Spring	Mod	2750	2	2	1	0
11	K-1492	2	Spring	Mod	1800	19	20	9	9
12	K-1493	1	Spring	Mod	1600	12	12	6	6
13	B-402	1	Spring	Mod	2000	6	5	4	1
14	B-407	1	Spring	Mod	1000	11	8	7	4
15	B-444	8	Spring	Mod	6189	4	3	3	2
16	B-447	1	Spring	Mod	2500	2	2	2	0
17	B-453	1	Spring	Mod	7000	19	13	9	12
18	B-457	4	Spring	Mod	7800	2	2	2	2
19	K-1448	1	Straight	Mod	2809	8	8	2	0
20	K-1450	1	Straight	Mod	2500	12	12	6	7
21	K-1449	2	Straight	Mod	3700	3	3	1	0
22	K-1451	1	Straight	Mod	3000	6	6	2	1
23	K-1464	8	Straight	Mod	1500	4	4	1	0
24	K-1465	1	Straight	Mod	1509	11	11	5	2
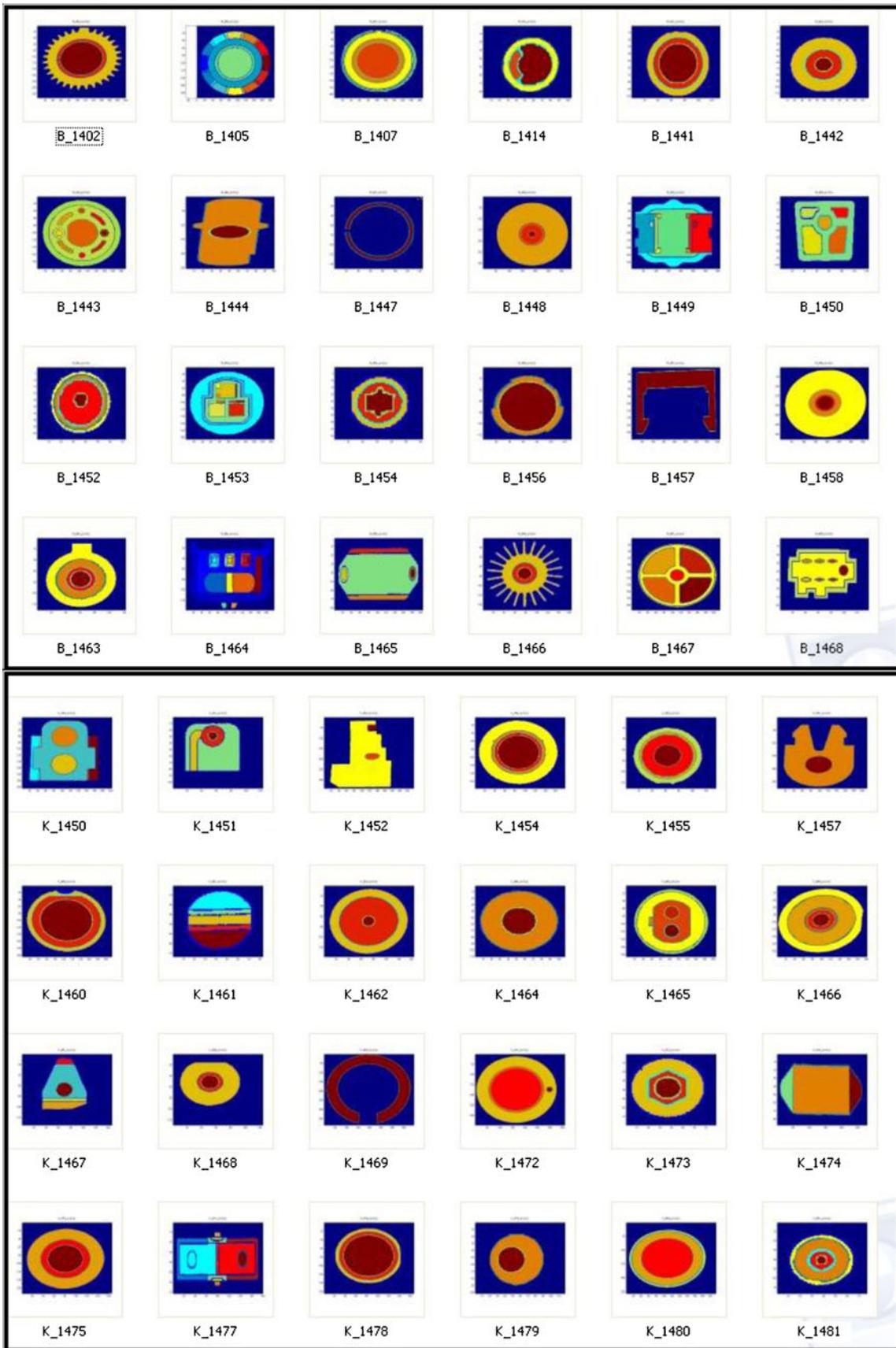
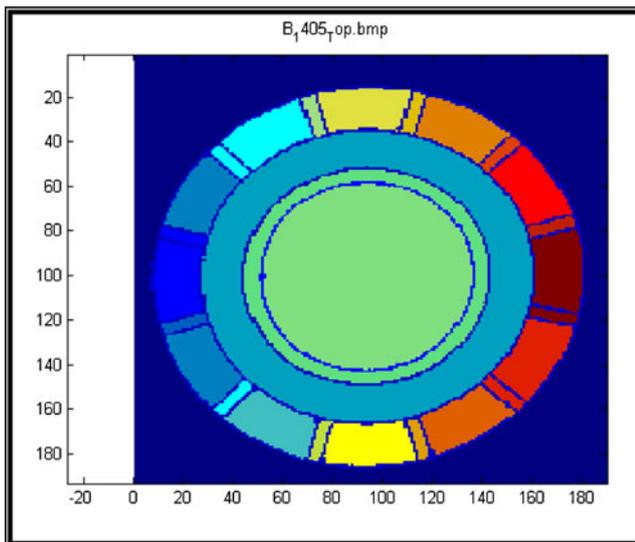**Fig. 7** Scanned images of different parts with boundaries detected

Fig. 8 An example of symmetric part

are similar based upon all metrics, then they belong to the same cluster. Symmetry indirectly measures complexity and cost.

Additionally, the geometry-related binary variable *NonCir* is used to indicate the circularity of boundaries. Non-circular boundaries are typically more difficult to machine and have a higher cost than circular ones. *NonCir* is calculated from the regional descriptor *eccentricity*. Boundaries with eccentricities above 0.4 are regarded as non-circular. This is a simple way to separate circular boundaries from more complex non-circular ones.

Another variable *washer* is also defined as binary. If the number of circular boundaries in a part is less than two, this variable is defined as a washer. Otherwise, it indicates a more complex shape. This variable plays a role in the
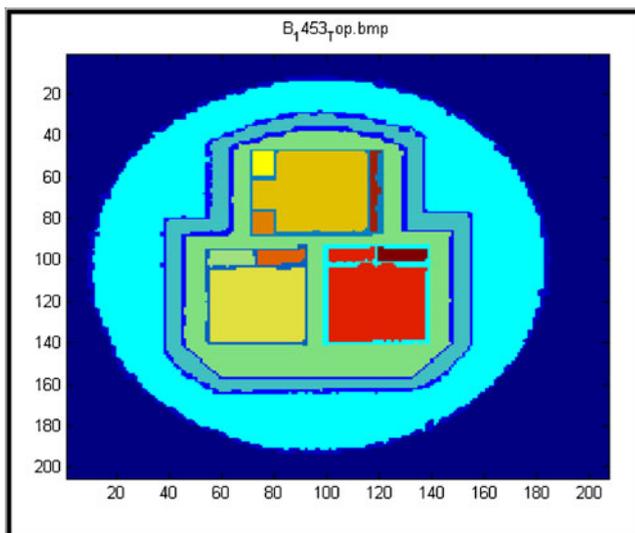

Fig. 9 An example of non-symmetric part

regression model for mold types. Most molds for washers are the design type of straight.

4.3 Analysis

4.3.1 Data partitions

The mixed dataset of 83 molds is partitioned into three distinct homogeneous datasets of mold types and mold design. The reason is twofold. First, the variance of residuals is reduced because the data have a smaller range. Second, outliers are easier to identify within the respective partitions.

The first partition is by mold type. A Kruskal–Wallis test is used to check the normality and equal variance of the partition. The p value of 0.001 for this test confirms that mold type can, in fact, be used to partition the dataset. After partitioning based on mold type, we use the modular partition for further tests because this subset is most similar to what we are estimating. To determine whether mold design is also a good partition, we perform a Tukey multiple comparisons tests. The individual confidence level is 98.07 %. The confidence interval did not include zero. Therefore, the result also indicates that the design cam is different from either spring or straight.

The difference between design spring and design straight was less clear. Thus we performed a t test to determine whether or not these are indeed good partitions. A p value of 0.12 is obtained. This is an indication that the populations were indeed different. Thus, we further partition the dataset by design. Based on the partitioned datasets, we can construct two regression models, as discussed in the next section. One is for the mold type of modular and the design type of straight. The other is for the mold type of modular and design type of spring.

4.3.2 Regression models of the partitioned datasets

Here we model two of the partitioned datasets. The first partition is of mold type modular and mold design straight. The second partitioned dataset is of mold type modular and mold design spring. Mold type modular and mold design cam are not modeled because the dataset only includes four observations.

1. Mold type modular and design type straight

   This partitioned dataset has 31 observations and is considered to be large enough to build a regression model with statistical significance. The first step in this process is the removal of outliers. Of the 31 observations, six outliers were identified. The six outliers had significant contributing cost factors related to some extremely complex geometry that is not representative for the overall samples.

Table 2 Comparisons of predicted and actual costs (mold type is *modular* and mold design is *straight*)

Predict	Lower	Upper	APE	PI	Name	Cost
1,523	222	2,824	45.78 %	2,602	K-1448	2,809
2,868	1,323	4,414	4.40 %	3,091	K-1451	3,000
1,388	−715	3,490	8.02 %	4,205	K-1465	1,509
2,429	754	4,103	10.41 %	3,349	K-1472	2,200
1,756	356	3,157	26.83 %	2,801	K-1458	2,400
2,146	304	3,988	12.95 %	3,684	K-1473	1,900
3,066	1,560	4,571	36.27 %	3,011	K-1480	2,250
1,702	62	3,342	13.47 %	3,280	B-441	1,500
2,886	1,291	4,481	15.44 %	3,190	B-452	2,500
3,255	1,618	4,893	41.52 %	3,275	B-448	2,300
MAPE			21.51 %			

The next step was variable selection. We removed some variables, such as the number of cavities (*NumCav*), because of the associated high p value, which does not appear to contribute to the predictive ability. Rather, we find that the quadratic version of this variable is significant and retain it in the model.

Our final model includes 25 observations with approximately 68 % of the variance in the data being explained by the model. The final fitted model is

$$Cost = 1981 + 3.04WaveClust - 744RegClust \\ + 1152NonCir - 13.2NumCav^2 + 111RegClust^2 \\ - 103NonCir^2 - 1707Washer - 25.7WaveClust^2$$

2. Mold type modular and mold design spring

The second partition includes 14 observations that are of type modular and design spring. We find one obvious outlier since a further investigation reveals that this mold was incorrectly coded, which was an unusual design type called double spring load. The next step is variable selection. We excluded *reg_sq*, which was the second-order version of *RegClust*, because it has a high

p value and does not contribute to the model's predictive ability. The linear term of another variable *NumBound* also highly correlates with *RegClust* and *WaveClust* and therefore was removed.

Our final model includes 13 observations with approximately 99 % of the variance explained by the model. The final fitted model is

$$Cost = 1737 + 731NumCav + 562WaveClust \\ - 653RegClust - 701NonCir + 20.8NumBound^2 \\ - 33.5WaveClust + 67.2NonCir^2$$

One important consideration of the above two regression models is that we are working with observed data instead of data from properly designed experiments. Therefore, we cannot explore factors in isolation. These variables are not strictly independent of each other. As a result, the derived regression models are used strictly for prediction instead of explanation.

4.4 Cross-fold validation

Here we validate the regression models using the cross-validation approach. Because of the relatively small sample sizes, the leave-one-out strategy is used. The leave-one-out strategy is to take one observation out of the dataset for validation and use the rest for modeling. This procedure is repeated for each observation in the dataset.

Using this strategy, we calculate several validation metrics from the two regression models as shown in Tables 2 and 3, respectively. Reading from left to right of the tables, the metrics include the cost prediction (*Predict*) based upon the regression model, the 95 % lower (*Lower*) and upper (*Upper*) prediction intervals, the absolute percentage error (*APE*), the prediction interval length (*PI*), as well as the name of the mold (*Name*) and its actual cost (*Cost*). The

Table 3 Comparisons of predicted and actual costs (mold type is *modular* and mold design is *spring*)

Predict	Lower	Upper	APE	PI	Name	Cost
2,873	2,194	3,554	19.71 %	1,360	K-1457	2,400
2,637	1,471	3,803	12.10 %	2,332	K-1474	3,000
3,349	2,586	4,113	16.28 %	1,527	K-1481	4,000
4,122	3,156	4,700	3.05 %	1,544	K-1482	4,000
4,123	3,157	5,088	3.08 %	1,931	K-1483	4,000
2,972	2,028	3,916	8.07 %	1,888	K-1487	2,750
1,927	370	3,484	20.44 %	3,114	K-1493	1,600
1,916	957	2,875	4.20 %	1,918	B-402	2,000
8,088	903	15,273	30.68 %	14,370	B-444	6,189
1,735	−4,062	7,537	3.61 %	11,599	K-1492	1,800
MAPE			12.12 %			

last row of Tables 2 and 3 is the mean absolute percentage error (*MAPE*), defined as

$$MAPE = |Predict - Cost|/Cost$$

From the values of MAPE, we can observe that the spring model is more accurate than the straight model. Considering the cost range of injection molds is very wide (can be as high as $60,000 and as low as $600) and the variability of the data is high, the MAPE of 21.51 % is actually remarkable. Recall that our models are based upon bids and not real manufacturing costs. Even if we had real manufacturing costs, it is well known that in a job shop environment some jobs may go well while others do not, which results in noisy data.

5 Conclusions and future extensions

In this paper, we presented a semi-automatic framework to estimate the cost of injection molds. It is believed to be difficult to fully automate this process because injection mold design itself is complex and human knowledge and experiences cannot be entirely replaced by databases. Therefore, an experienced bidder should always be included in the process.

It is demonstrated that images can be used as the common data format for automated mold cost estimation, from which geometric and topological features can be extracted automatically. A unique feature vector combines the knowledge of mold type, mold design, and geometry to construct regression models for cost estimation. Wavelet, regional, and topological descriptors are used to relate geometry similarity to complexity and cost.

Some unique variables that are able to relate geometry in 2D images to costs were identified, including the number of non-circular boundaries, the number of boundaries, and degrees of symmetry. The number of wavelet clusters and the number of regional clusters measure the degree of symmetry.

Our new hybrid approach combines analogy and mathematical methods. At the analogy-based data partitioning step, data are grouped into homogeneous datasets. This step is useful to reduce the variance of estimations. At the regression-modeling step, mathematical models give both mean estimates and prediction intervals. The information of variance provides a necessary tool for risk management.

Our approach, however, has several limitations. The first is that a human expert is still needed to identify the factors that are not captured by the model. The constructed model thus should only be used to assist, instead of replace, human bidders. The cost estimate from the model can serve two purposes: to provide a double check on the bidders' estimates and to assist in early design decisions. The second

limitation is that data preprocessing for images may require an extra amount of time. For instance, the removal of lines that are not associated with the geometry of the part must be done prior to our methodology, although converting various 2D or 3D design data formats to 2D images is easy. For our examples, all non-geometric line removal was done manually. This manual process took approximately 15 min per part. The automation of non-geometric line removal needs further investigation in future research.

The developed injection molding cost estimation process and framework can also be extended to other net shape manufacturing processes, such as casting, blow molding, transfer molding, compression molding, stamping, etc. The methods described in this paper form the basis to automate the tooling cost estimation for these processes.

References

1. El-Mehalawi M (1999) A geometric similarity case-based reasoning system for cost estimation in net-shape manufacturing. Ohio State University
2. Kwong DCK, Smith GF (1998) A computational system for process design of injection moulding: combining blackboard-based expert system and case-based reasoning approach. Int J Adv Manuf Technol 14(4):239–246
3. Wong TS (1998) Evaluation of a software package to estimate cost of an injection mold. University of Massachusetts, Lowell
4. Wang H, Ruan X-Y, Zhou XH (2003) Research on injection mould intelligent cost estimation system and key technologies. Int J Adv Manuf Technol 21(3):215–222
5. Bergmann R, Althoff K-D, Breen S, Göker M, Manago M, Traphöner R, Wess S (2004) Developing industrial case-based reasoning applications: the INRECA methodology. Springer, Berlin
6. Shehab E, Abdalla H (2002) An intelligent knowledge-based system for product cost modelling. Int J Adv Manuf Technol 19(1):49–65
7. Nagahanumaiah K, Mukherjee NP, Ravi B (2005) An integrated framework for die and mold cost estimation using design features and tooling parameters. Int J Adv Manuf Technol 26(9–10):1138–1149
8. Raviwongse R, Allada V (1997) Artificial neural network based model for computation of injection mould complexity. Int J Adv Manuf Technol 13(8):577–586
9. Shehab EM, Abdalla HS (2002) A design to cost system for innovative product development. Proc Inst Mech Eng B J Eng Manuf 216(7):999–1019
10. Chan SF, Law CK, Chan KK (2003) Computerised price quoting system for injection mould manufacture. J Mater Process Technol 139(1–3):212–218
11. Sapene C (2007) Cost analysis of plastic injection molds. Lulu.com
12. Latecki L, Lakamper R (2000) Shape similarity measure based on correspondence of visual parts. IEEE Trans Pattern Anal Mach Intell 22(10):1185–1190
13. Tangelder JWH, Veltkamp RC (2004) A survey of content based 3D shape retrieval methods. In: Shape Modeling Applications, 2004. Proceedings, pp. 145–156
14. Lou K, Ramani K, Prabhakar S (2004) Content-based three-dimensional engineering shape search. In: 20th International Conference on Data Engineering, 2004. Proceedings, pp. 754–765

15. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522
16. Loncaric S (1998) A survey of shape analysis techniques. Pattern Recogn 31(8):983–1001
17. Iyer N, Jayanti S, Lou K, Kalyanaraman Y, Ramani K (2005) Three-dimensional shape searching: state-of-the-art review and future trends. Computer Aided Des 37(5):509–530
18. Cardone A, Gupta RK, Karnik M (2003) A survey of shape similarity assessment algorithms for product design and manufacturing applications. J Comput Inf Sci Eng 3:109–118
19. Bustos B, Keim DA, Saupe D, Schreck T, Vranić DV (2005) Feature-based similarity search in 3D object databases. ACM Comput Surv 37(4):345–387
20. Walker JS (1999) A primer on wavelets and their scientific applications. CRC, Boca Raton
21. Matthews I, Xiao J, Baker S (2007) 2D vs. 3D deformable face models: representational power, construction, and real-time fitting. Int J Comput Vis 75(1):93–113
22. Kluz J, D. & P. M. A. National Tool, and National Tooling and Machining Association (U.S.) (1981) Moldmaking & die cast dies for metalworking trainees: machine tool operators, machinists, toolmakers, diemakers, moldmakers, special machine builders. National Tooling & Machining Association, Washington
23. Gonzalez RC, Woods RE, Eddins SL (2004) Digital image processing using MATLAB. Pearson Prentice Hall, Upper Saddle River